

MSC

2.º
CICLO

FCUP
ANO

U.PORTO

Modelo de rating

António Miguel Arantes da Silva

FC

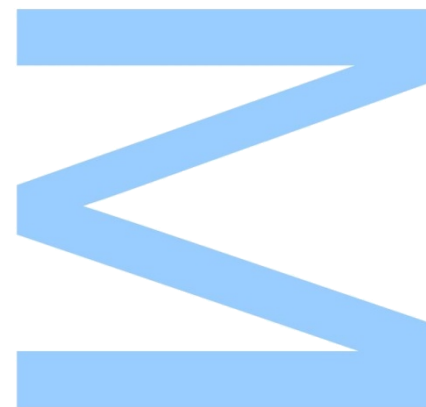
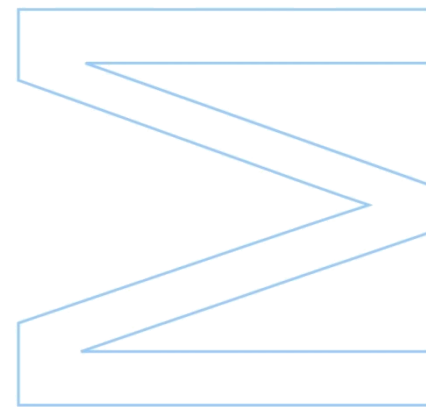
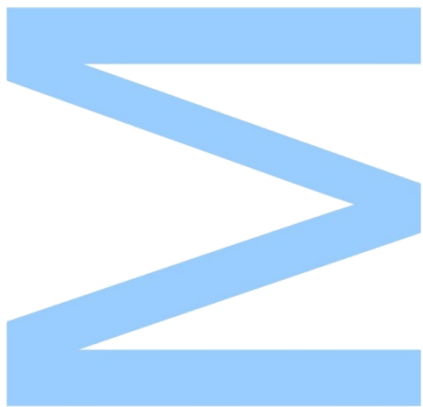


Modelo de Rating

António Miguel Arantes da Silva

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Engenharia Matemática

2016



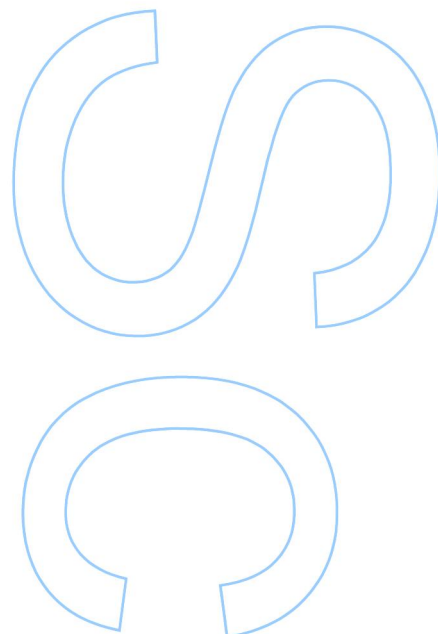
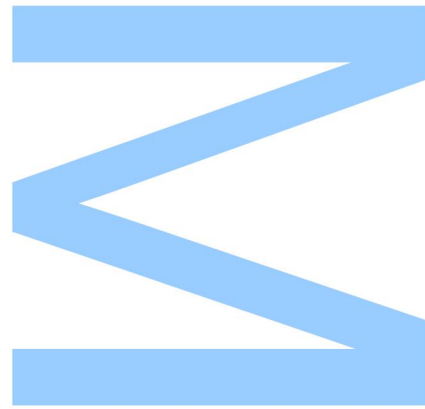
Modelo de Rating

António Miguel Arantes da Silva

Engenharia Matemática
Departamento de Matemática
2016

Orientador

Alberto Pinto, Professor Catedrático, FCUP

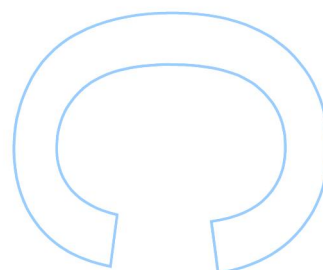
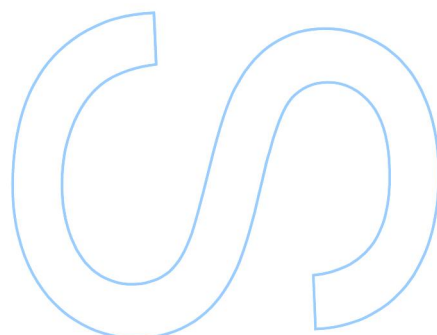
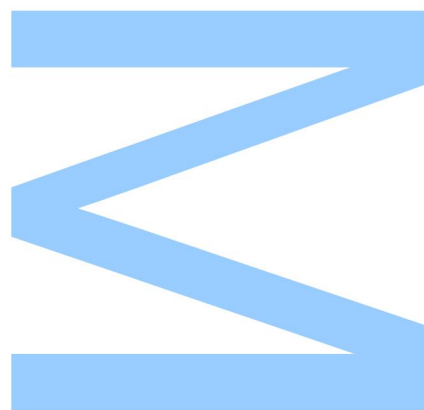




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____



Agradecimentos

Em primeiro lugar, ao meu orientador, Professor Doutor Alberto Pinto, por toda a ajuda prestada na elaboração deste trabalho, sem a qual não teria conseguido alcançar este objetivo.

À minha família, em especial à minha mulher, por toda a paciência demonstrada para comigo e por todo o apoio que me deu sempre! Quero também pedir desculpa pela ausência “física e psicológica” durante o tempo de conclusão deste mestrado. Quero agradecer todo o carinho e apoio dado ao longo destes anos.

Ao meu filho que, mesmo quando eu não lhe podia dar a atenção merecida, nunca me falhou com imenso carinho, três beijos doces e a mais bela frase que poderia ouvir ”Pai, pai, pai!” que gritou inúmeras vezes para me chamar a atenção.

À minha filha que, mesmo sem entender as ausências do pai, sempre que me vê, faz uns sorrisos enormes de felicidade. Além disto, conseguiu dar-me umas belas noites mal dormidas.

Como, em tão pouco tempo, vocês cresceram.

Graças a vocês consegui cumprir um dos objetivos da minha vida.

Obrigado!

Motivação

A minha motivação poderia ser resumida num único provérbio: "A necessidade aguça o engenho."

E essa foi, durante muito tempo, a motivação para terminar este documento: a necessidade de ter um modelo de rating que respondesse às necessidades da instituição financeira onde trabalhava e onde estava totalmente enquadrado no departamento de gestão de riscos.

Contudo, o tempo avança e nós também avançamos. Acabei por mudar de emprego e a minha motivação para terminar a tese acabou por esgotar-se, tendo-se protelando no tempo a finalização da mesma, levando a que quase desistisse de a terminar.

Assim, a tese ficou parada e na vida "visão sem ação é sonho e ação sem visão é pesadelo", tornou-se imperativo e uma questão de orgulho pessoal dar à empresa que me apoiou durante 5 anos um modelo melhor do que o deixado, detalhado e explicado de forma a que qualquer novo matemático contratado, ou a trabalhar no setor financeiro e com vontade de aprender, pudesse utilizar este documento como manual para elaboração de um modelo robusto de rating.

Na minha opinião, o departamento de gestão de riscos deve assumir um papel cada vez mais preponderante na análise estratégica das instituições financeiras, sendo este um elemento fundamental para os órgãos de gestão, fazendo com que o risco de crédito assuma uma posição de excelência no atual sistema financeiro.

O rating é hoje visto como um instrumento de informação para todos os departamentos de uma instituição de crédito, pois pode ser definido como um cálculo de credibilidade de um tomador de crédito cumprir total ou parcialmente com as obrigações assumidas. Portanto, ser capaz de criar um modelo de rating torna-se uma mais-valia para qualquer analista de risco de crédito.

Resumo

Devido à crise de 2008, o regulador tem intervido mais ativamente junto do setor financeiro, através de legislação para controlo de riscos das instituições cada vez mais exigente (ver [5]).

Esta dissertação foi desenvolvida a partir de uma base de dados reais, de uma instituição financeira, construída especificamente para o efeito; é composta por um conjunto de empresas dos mais diversos setores económicos e está aqui representada uma parte substancial do universo das micro empresas, pequenas e médias empresa em Portugal, contendo, ainda, uma pequena amostra das grandes empresas nacionais. No total temos em média cerca de 40 mil empresas nos últimos 3 anos, em comparativo com o relatório publicado pela Informa D&B (ver [10]) existem em Portugal cerca de 456 mil empresas. Esta base de dados contém informações económicas e financeiras destas empresas, que funcionam nos mais variados setores de atividade nacionais.

Nesta dissertação, propõem-se a construção de um modelo de rating capaz de satisfazer as necessidades de prever a qualidade creditícia de um cliente e de responder às expectativas da instituição financeira.

A metodologia utilizada limitou-se a seguir o usual neste tipo de modelos no sector bancário, ou seja, após a análise de cada uma das variáveis individualmente e no conjunto, utilizou-se a regressão univariada e multivariada para escolher as variáveis explicativas do incumprimento, além do que, a regressão permite-nos compreender comportamento de uma variável dependente (incumprimento), em função de uma ou várias variáveis independentes ou explicativas.

Serão elaborados vários modelos, nos primeiros utilizar-se-à todas as variáveis exceto as excluídas por razões de cariz económico. Nos segundos modelos, irão ficar apenas variáveis cujas correlações são inferiores a 0,9. Nos terceiros modelos, irão ficar apenas variáveis cujas correlações são inferiores a 0,7.

Será feita a comparação dos modelos utilizando o AIC (A Information Criterion) de Akaike e a curva de ROC (Receiver Operating Characteristic).

Será elaborado um quadro comparativo de cada modelo obtido. E espera-se obter um modelo, adequado e parcimonioso, com capacidade discriminante.

Palavras chave: Modelo de *rating*, regressão logística, classificação de empresas, modelo de *scoring*.

Abstract

Because of the 2008 crisis, the regulator has intervened more actively with the financial sector, through legislation for risk control to increasingly demanding financial institutions (see [5]).

This dissertation was developed from a real database from financial a institution, built specifically for this purpose; it consists in a set of companies from different economic sectors and represent a substantial part of the universe of micro enterprises, small and medium-sized enterprises in Portugal, containing also a small sample of big national enterprises. In total we have about 40 thousand enterprises in average for the three years, in comparison to the published by Informa D&B (see [10]) are now in Portugal around 456 thousand enterprises. This database contains economic and financial information of these enterprises, which operate in various national sectors.

In this dissertation, we propose the construction of a rating model able to meet the needs of predicting the creditworthiness of a customer and to meet the expectations of the financial institution.

The methodology used was limited to following the usual in this type of models in the banking sector, that is, after the analysis of each of the variables individually and together, we used univariate and multivariate regression to choose the explanatory variables of the model, besides this regression allows us to understand the behavior of a dependent variable (default), due to one or more independent or explanatory variables.

Different models will be developed, the first one will use all variables except be excluded for economic reasons. In the second model, there will be only variables whose correlations are less than 0.9. In third model, will be only variables whose correlations are less than 0.7.

There will be a comparison of the models using the AIC (A Information Criterion) of Akaike and the ROC (Receiver Operating Characteristic) curve.

They will be a comparative table of each model. And it's expected obtain a model, adequate and parsimonious, with discriminant capacity.

Keywords: Rating model, logistic regression, business classification, scoring model.

Índice

Agradecimentos	I
Motivação	III
Resumo	V
Abstract	VII
1 Introdução	7
2 Análise do ficheiro de dados	9
2.1 Introdução	9
2.2 Análise das variáveis	9
3 Modelo de rating	17
3.1 Introdução	17
3.2 O modelo de regressão logística	18
3.2.1 Caso univariado	19
3.2.2 Caso multivariado	22
3.3 Análise preditiva do modelo	23
3.3.1 Teste de Wald	23
3.3.2 Critério de Informação de Akaike	25
3.3.3 Curva de ROC	25
3.4 Amostra de treino e amostra de teste	27
3.5 Escalões de rating	29
3.6 Variáveis qualitativas	31
4 Aplicação do modelo de rating	32
4.1 Algoritmo aplicado para selecção do modelo de rating	32
4.2 Aplicação do modelo de regressão logística aos dados	33
4.2.1 Modelo construídos a partir da base de dados completa	36
4.2.2 Modelos construídos sem correlações superiores a 0,90	41

4.2.3	Modelos construidos sem correlações superiores a 0,70	45
5	Conclusão	50
6	Desenvolvimentos futuros	54
Appendix A	Tabela de correlações de todas as variáveis contabilísticas	57

Lista de figuras

2.1	Distribuição de clientes por ano: incumprimento (D) e cumprimento (\tilde{D})	10
2.2	Distribuição das empresas por CAE	12
2.3	Distribuição das empresas por categoria	14
3.1	Gráfico da função logística	20
3.2	Curva de ROC	27
4.1	Coeficientes do modelo das micro empresas	36
4.2	Análise da curva de ROC do modelo para as micro empresas	37
4.3	Análise da curva de ROC do modelo para as micro empresas	37
4.4	Coeficientes do modelo das pequenas, médias e grandes empresas	39
4.5	Análise da curva de ROC do modelo para as micro empresas	39
4.6	Análise da curva de ROC do modelo para as micro empresas	40
4.7	Coeficientes do modelo das micro empresas	42
4.8	Análise da curva de ROC do modelo para as micro empresas	42
4.9	Coeficientes do modelo das pequenas, médias e grandes empresas	44
4.10	Análise da curva de ROC do modelo para as micro empresas	44
4.11	Coeficientes do modelo das micro empresas	46
4.12	Análise da curva de ROC do modelo para as micro empresas	46
4.13	Coeficientes do modelo das pequenas, médias e grandes empresas	48
4.14	Análise da curva de ROC do modelo para as micro empresas	48

Lista de tabelas

2.1	Distribuição das empresas por ano	9
2.2	Agregações dos CAE	11
2.3	Distribuição das empresas por CAE	12
2.4	Categoria de uma empresa - critérios	13
2.5	Distribuição das empresas por categoria	13
2.6	Correlação superior de 0.9 - parte I.	16
2.7	Correlação superior a 0.9 - parte II.	16
3.1	Escalões de rating reconhecidos pelo Banco de Portugal para avaliação externa .	30
3.2	Escalões de rating reconhecidos pelo Banco de Portugal para avaliação externa .	30
3.3	Escalões de rating utilizados	30
4.1	Distribuição das empresas pelas amostras de treino e teste, por categoria	34
4.2	Distribuição das empresas pelas amostras de treino e teste, por CAE	34
4.3	AIC do modelo de regressão logística univariado	35
4.4	AIC e ROC dos modelos com as variáveis sem restrições de correlações	38
4.5	Escalões de rating para as micro empresas	38
4.6	AIC e ROC dos modelos com as variáveis sem restrições de correlações	40
4.7	Escalões de rating para as pequenas, médias e grandes empresas	41
4.8	AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.9 . .	43
4.9	Escalões de rating para as micro empresas	43
4.10	AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.9 . .	45
4.11	Escalões de rating para as pequenas, médias e grandes empresas	45
4.12	AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.7 . .	47
4.13	Escalões de rating para as micro empresas	47
4.14	AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.7 . .	49
4.15	Escalões de rating para as pequenas, médias e grandes empresas	49
5.1	Análise comparativa dos modelos para as micro empresas	51
5.2	Variáveis explicativas dos modelos para as micro empresas	51
5.3	Análise comparativa dos modelos para as pequenas, médias e grandes empresas .	52

5.4 Variáveis explicativas dos modelos para as pequenas, médias e grandes empresas 53

Capítulo 1

Introdução

O crédito concedido às empresas é parte fundamental não só para o seu funcionamento, mas fundamentalmente para a economia do país, facilitando transações de produtos, bens e serviços.

Associado ao crédito, existe o fator risco. Para administrar e controlar este risco (de crédito), utilizam-se diversas metodologias de mensuração de exposição e mecanismos de gestão. Destes métodos, destacam-se os seguintes como técnicas utilizadas para a construção de modelos de risco de crédito: programação linear, análise discriminante linear e quadrático, regressão logística, árvores de decisão, redes neurais, método dos K-vizinhos mais próximos e máquinas de suporte vetorial e suas variações.

Devido à crise de 2008, o regulador tem intervido mais ativamente junto do setor financeiro, através de legislação para controlo de riscos das instituições cada vez mais exigente (ver [5]). Torna-se, por isso, cada vez mais relevante nesta atividade o controlo do principal risco que lhe está associado, o já referido risco de crédito - que consiste na probabilidade de ocorrência de impactos negativos nos resultados ou no capital, devido à incapacidade de uma contraparte cumprir os seus compromissos financeiros perante a instituição.

A capacidade de discriminar bons e maus clientes tornou-se, assim, um fator decisivo para o sucesso das instituições financeiras.

Neste trabalho, vai ser analisado um conjunto de dados relativamente à capacidade creditícia de algumas empresas nacionais, com o objetivo de estimar quais as empresas cujos contratos celebrados nos empréstimos bancários não serão cumpridos.

De forma a atingir o objetivo proposto, vai-se começar por analisar uma base de dados económicos e financeiros fornecida por uma instituição financeira de relevância nacional. Nesta lista estão contidas informações de milhares empresas nacionais, desde micro a grandes empresas, que são recolhidas para analisar do risco de crédito de cada empresa.

Após esta análise, vai-se apresentar a construção de um modelo, com recurso à regressão logística, com o objetivo de categorizar estas mesmas empresas, relacionando o cumprimento e incumprimento do pagamento do crédito concedido com algumas das restantes informações presentes na base de dados. Através desta categorização, pretende-se criar uma previsão para

avaliar qual o fator de risco de um novo indivíduo com o menor valor de erro possível.

Posteriormente, para aproximar esta informação ao que é aplicado na realidade nos métodos mais discriminantes das empresas, vão-se elaborar modelos de rating.

Capítulo 2

Análise do ficheiro de dados

2.1 Introdução

O objetivo deste trabalho é a construção de um modelo de *rating* para ser aplicado numa base de dados económicos e financeiros de empresas de vários setores de atividade nacionais e de várias categorias (desde micro a grandes empresas) que foi gentilmente fornecida por uma instituição financeira. Ao documento original foram feitas algumas alterações que se descrevem a seguir para garantir a confidencialidade das informações nele contidas.

Nesta secção, para além da apresentação das informações presentes na base de dados, vai ser ainda feita uma análise às variáveis contidas na mesma para uma melhor compreensão destas e para ajudar a perceber os resultados obtidos nos métodos que serão apresentados ao longo deste trabalho.

2.2 Análise das variáveis

Para a análise do risco de crédito de uma empresa, são recolhidas várias informações que contribuem com maior ou menor peso para este risco. Na base de dados fornecida, estão contidas as seguintes informações: Classificação Portuguesa de Atividades Económicas da empresa (CAE), a categoria desta e as contas no sistema de normalização contabilística (SNC).

Foram gentilmente cedidos 3 anos de informação estatística; a distribuição do número de empresas por ano está representada na tabela seguinte:

Ano	Incumprimento	Em cumprimento
2011	1914	38334
2012	1196	37206
2013	993	38739

Table 2.1: Distribuição das empresas por ano

No gráfico seguinte estão representados estes valores, para uma melhor visualização da sua distribuição:

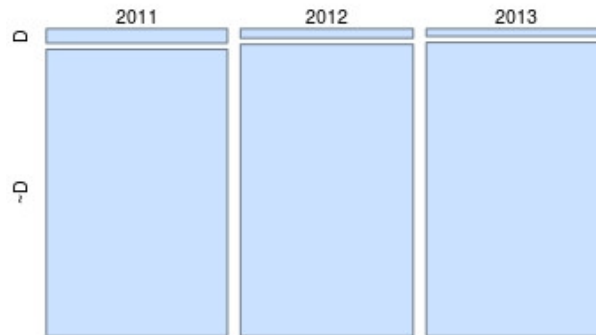


Figure 2.1: Distribuição de clientes por ano: incumprimento (D) e cumprimento (\tilde{D})

Foram fornecidos os dados de 118382 empresas.

As variáveis nominais CAE, concelho e categoria da empresa estão devidamente identificadas, como tal é possível fazer-lhes uma análise mais detalhada.

A Classificação Portuguesa de Atividades Económicas (CAE), cuja primeira versão remonta ao ano de 1953, é elaborada pelo Instituto Nacional de Estatística (INE) (ver [11]) e tem como objetivo a classificação e agrupamento das atividades económicas (produção, emprego, energia, investimento, etc) em unidades estatísticas de bens e serviços.

Utilizou-se a instrução 5/2011 do Banco de Portugal disponível em [2] para agregar os CAE, considerando-se as seguintes agregações:

Código CAE	Sectores de atividade económica
A	- Agricultura, produção animal, caça, floresta e pesca
B	- Indústrias extrativas
C	- Indústrias transformadoras
D,E	- Eletricidade, gás, água
F	- Construção
G	- Comércio e reparações
H	- Transportes e armazenagem
I	- Alojamento, restauração e similares
J	- Atividades de informação e de comunicação
K	- Atividades financeiras e de seguros
L	- Atividades imobiliárias
M, N	- Outros serviços empresariais
O	- Administração pública
P, Q	- Educação, saúde e apoio social
R, S, U	- Outras atividades

Table 2.2: Agregações dos CAE

Como na informação prestada não existiam empresas do sector O e U, os mesmos foram retirados das análises, visto a sua frequência ser 0.

De seguida, apresenta-se a tabela e o gráfico com a distribuição das empresas (presentes no ficheiro) por CAE, conforme a instrução do Banco de Portugal indicada anteriormente:

Código CAE	2011	2012	2013
A	441	570	767
B	158	147	132
C	7654	7392	7680
D,E	170	180	188
F	4298	3730	3541
G	14166	13653	14292
H	1807	1768	1869
I	2613	2397	2300
J	765	759	801
K	109	166	266
L	701	594	519
M, N	4598	4391	4457
P, Q	2005	1978	2113
R, S	763	727	807

Table 2.3: Distribuição das empresas por CAE

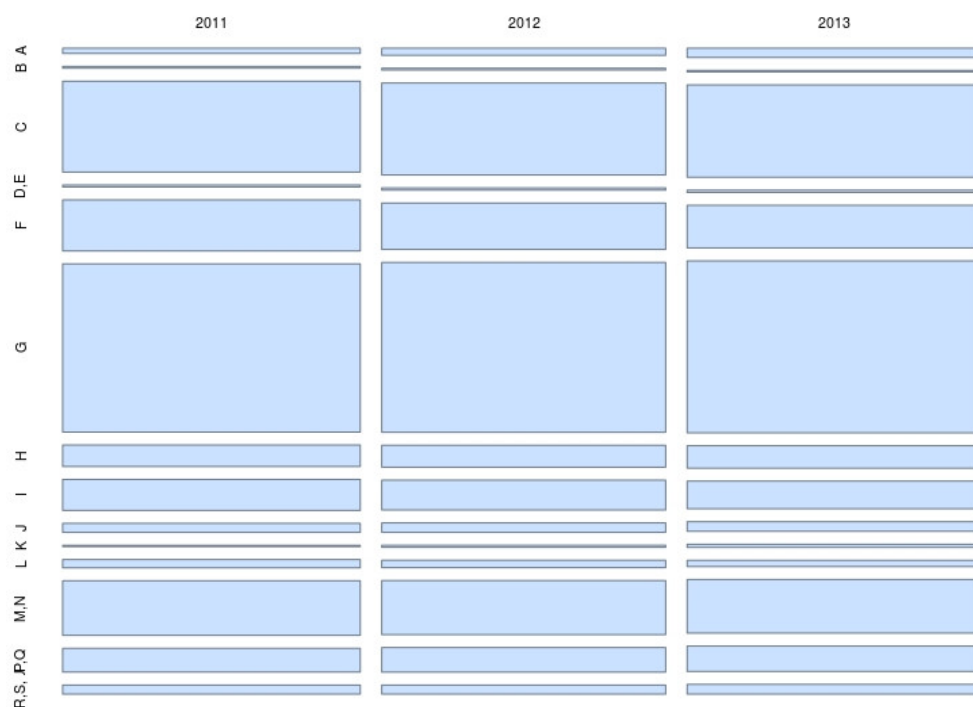


Figure 2.2: Distribuição das empresas por CAE

A categoria da empresa - micro, pequena, média ou grande - foi determinada de acordo com o Decreto-Lei n.º 372/2007. Esta categoria é encontrada de acordo com os 3 seguintes critérios:

- Efetivos: unidades de trabalho por ano;
- Volume de negócios anual;
- Balanço total anual.

O artigo 42º do Código do Imposto sobre o Valor Acrescentado (CIVA) define o volume de negócios de uma empresa como sendo o total de vendas realizadas durante um determinado período de tempo, com a exceção do IVA. O balanço de uma empresa será definido à frente, neste capítulo.

Tendo por base estes critérios, as empresas são categorizadas do seguinte modo:

Categoria	Efetivos	Volume de negócios	ou	Balanço total
Micro	<10	≤2 milhões de euros		≤2 milhões de euros
Pequena	<50	≤10 milhões de euros		≤10 milhões de euros
Média	<250	≤50 milhões de euros		≤43 milhões de euros
Grande	≥250	>50 milhões de euros		>43 milhões de euros

Table 2.4: Categoria de uma empresa - critérios

A título de exemplo, uma média empresa é definida como uma empresa que emprega menos de 250 pessoas e cujo volume de negócios anual não excede os 50 milhões de euros ou balanço total anual não excede 43 milhões de euros.

A seguir, apresenta-se a tabela da distribuição e o gráfico, por tamanho de empresa, de acordo com os critérios anteriormente mencionados:

Categoria	2011	2012	2013
Micro	24750	24033	25464
Pequena	12660	11729	11709
Média	2489	2403	2333
Grande	249	237	266

Table 2.5: Distribuição das empresas por categoria

Ontem a falar com o Alberto, pensamos que, de forma a melhorar o anonimato, se poderia retirar as tabelas 2.6, 2.7 e 2.10 e o texto respetivo, assim como as colunas dos coeficientes de estimação e erros padrão da regressão logística (ficando a identificação da variável, o valor de z e de p ; já agora, a ordenada na origem também não faz falta).



Figure 2.3: Distribuição das empresas por categoria

Em relação às informações relativas às contas SNC de cada empresa fornecidas pela instituição financeira, mudou-se o nome destas variáveis para VAR seguido de um número para garantir a confidencialidade dos dados, isto é, as variáveis passaram a chamar-se VAR01, VAR02, ..., VAR72. Antes da alteração do nome das variáveis, era possível diferenciar as seguintes grandes rubricas:

- Balanço
O balanço é a relação entre o ativo e o passivo, ou seja, permite estimar o valor da empresa numa determinada data.
- Demonstração de resultados
Mostra o resultado líquido positivo (lucro) ou negativo (prejuízo, perda) alcançado num determinado período.

E alguns rácios contabilísticos, usuais entre os analistas de crédito, calculados com estas mesmas contas, dos quais se destacam os seguintes:

- Rácio de Liquidez Corrente
Mede a solvência da empresa. Indica, a partir dos ativos correntes, a capacidade de pagar as dívidas (exigibilidades) correntes.
- Rácio de Liquidez Imediata
É uma medida mais próxima da liquidez efetiva da empresa porque elucida a capacidade dos seus ativos de maior liquidez para assegurarem a cobertura do passivo corrente ou exigível de curto prazo.
- Rotação Média de Existências
Mede o número de vezes que o inventário médio é completamente renovado (pela venda ou reposição) durante um período de análise.

- Prazo Médio de Pagamentos e Recebimentos

Este rácio expressa o número de dias que a empresa leva a pagar (ou receber) os seus débitos (créditos).

- Margem de Lucro Sobre as Vendas

Este rácio mede o lucro da empresa por cada euro de vendas.

As definições apresentadas anteriormente podem ser consultadas em [14].

Desta forma, não é possível elaborar uma análise mais detalhada de cada uma destas variáveis, optando-se por uma análise de correlação, análise de médias, extremos e quartis. Contudo, para manter o anonimado dos dados recebidos da instituição financeira o mesmo foi retirado desta dissertação.

Note-se que, tendo em conta as alterações feitas ao ficheiro de dados para manter a confidencialidade dos mesmos, não vai ser, igualmente, possível uma análise das variáveis do ponto de vista económico, o que irá condicionar os resultados obtidos nos modelos.

Contudo, foi feita uma análise da correlação entre as variáveis que se querem explicativas, considerando os dados de todos os anos (desde 2011 até 2013).

A correlação entre duas variáveis aleatórias indica a força e a direção do relacionamento linear entre elas. No uso estatístico geral, a correlação refere-se à medida da relação entre duas variáveis, embora correlação não implique causalidade. Assim, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados.

Para os dados em estudo, foi utilizado o coeficiente de correlação de Pearson, ρ , para a análise referida, que é obtido através do quociente entre a covariância entre as duas variáveis e o produto dos seus desvio-padrão. O valor deste coeficiente varia entre -1 e 1; dependendo do valor de ρ , a correlação entre as variáveis tem designações diferentes (ver [13]):

- $\rho = 1$: a correlação diz-se perfeita positiva;
- $\rho = -1$: a correlação diz-se perfeita negativa;
- $|\rho| > 0,90$: a correlação diz-se muito forte;
- $0,70 < |\rho| < 0,90$: a correlação diz-se forte;
- $0,50 < |\rho| < 0,70$: a correlação diz-se moderada;
- $0,30 < |\rho| < 0,50$: a correlação diz-se fraca;
- $0 < |\rho| < 0,30$: a correlação diz-se muito fraca;
- $\rho = 0$: a correlação diz-se nula, pelo que não existe uma relação linear entre as variáveis.

Em relação à primeira análise de correlação referida, apresentam-se a seguir os valores das correlações mais significativas:

Variável	VAR01	VAR03	VAR04	VAR05	VAR07	VAR08	VAR09	VAR10	VAR17	VAR20	VAR21
VAR01	1	0.55	0.42	0.64	0.59	0.28	0.24	0.24	0.73	0.43	0.47
VAR03	0.55	1	0.95	0.45	0.41	0.88	0.85	0.85	0.63	0.22	0.3
VAR04	0.42	0.95	1	0.3	0.25	0.97	0.94	0.94	0.52	0.04	0.11
VAR05	0.64	0.45	0.3	1	0.9	0.08	0.04	0.04	0.83	0.72	0.78
VAR07	0.59	0.41	0.25	0.9	1	0	0.03	0.03	0.7	0.74	0.77
VAR08	0.28	0.88	0.97	0.08	0	1	0.98	0.98	0.36	0.15	0.08
VAR09	0.24	0.85	0.94	0.04	0.03	0.98	1	1	0.3	0.16	0.1
VAR10	0.24	0.85	0.94	0.04	0.03	0.98	1	1	0.3	0.16	0.1
VAR17	0.73	0.63	0.52	0.83	0.7	0.36	0.3	0.3	1	0.49	0.58
VAR20	0.43	0.22	0.04	0.72	0.74	0.15	0.16	0.16	0.49	1	0.96
VAR21	0.47	0.3	0.11	0.78	0.77	0.08	0.1	0.1	0.58	0.96	1
VAR24	0.69	0.58	0.46	0.88	0.77	0.28	0.23	0.23	0.94	0.54	0.63
VAR25	0.67	0.53	0.38	0.93	0.84	0.17	0.13	0.13	0.9	0.74	0.83
VAR27	0.99	0.59	0.46	0.69	0.63	0.32	0.27	0.27	0.78	0.46	0.51
VAR30	0.01	0.02	0.03	0.02	0.02	0.04	0.03	0.03	0.02	0.04	0.04
VAR31	0.02	0.02	0.02	0.04	0.05	0.04	0.04	0.04	0.01	0.04	0.04
VAR32	0	0.03	0.03	0.03	0.03	0.04	0.05	0.05	0	0.03	0.03
VAR33	0	0.02	0.02	0.01	0.01	0.03	0.02	0.02	0.02	0.03	0.03
VAR38	0.09	0	0.01	0.05	0.05	0.02	0.02	0.02	0.02	0.08	0.08
VAR39	0.09	0	0.01	0.05	0.05	0.03	0.03	0.03	0.02	0.08	0.08
VAR45	0.24	0.85	0.94	0.04	0.03	0.98	1	1	0.3	0.16	0.1

Table 2.6: Correlação superior de 0.9 - parte I.

Variável	VAR24	VAR25	VAR27	VAR30	VAR31	VAR32	VAR33	VAR38	VAR39	VAR45
VAR01	0.69	0.67	0.99	0.01	0.02	0	0	0.09	0.09	0.24
VAR03	0.58	0.53	0.59	0.02	0.02	0.03	0.02	0	0	0.85
VAR04	0.46	0.38	0.46	0.03	0.02	0.03	0.02	0.01	0.01	0.94
VAR05	0.88	0.93	0.69	0.02	0.04	0.03	0.01	0.05	0.05	0.04
VAR07	0.77	0.84	0.63	0.02	0.05	0.03	0.01	0.05	0.05	0.03
VAR08	0.28	0.17	0.32	0.04	0.04	0.04	0.03	0.02	0.03	0.98
VAR09	0.23	0.13	0.27	0.03	0.04	0.05	0.02	0.02	0.03	1
VAR10	0.23	0.13	0.27	0.03	0.04	0.05	0.02	0.02	0.03	1
VAR17	0.94	0.9	0.78	0.02	0.01	0	0.02	0.02	0.02	0.3
VAR20	0.54	0.74	0.46	0.04	0.04	0.03	0.03	0.08	0.08	0.16
VAR21	0.63	0.83	0.51	0.04	0.04	0.03	0.03	0.08	0.08	0.1
VAR24	1	0.96	0.75	0.03	0.04	0.02	0.02	0.02	0.02	0.23
VAR25	0.96	1	0.73	0.04	0.04	0.03	0.03	0.05	0.05	0.13
VAR27	0.75	0.73	1	0.01	0.02	0	0	0.08	0.08	0.27
VAR30	0.03	0.04	0.01	1	0.37	0.38	0.91	0.18	0.18	0.03
VAR31	0.04	0.04	0.02	0.37	1	0.96	0.1	0.03	0.03	0.04
VAR32	0.02	0.03	0	0.38	0.96	1	0.09	0.05	0.05	0.05
VAR33	0.02	0.03	0	0.91	0.1	0.09	1	0.24	0.24	0.02
VAR38	0.02	0.05	0.08	0.18	0.03	0.05	0.24	1	0.99	0.02
VAR39	0.02	0.05	0.08	0.18	0.03	0.05	0.24	0.99	1	0.03
VAR45	0.23	0.13	0.27	0.03	0.04	0.05	0.02	0.02	0.03	1

Table 2.7: Correlação superior a 0.9 - parte II.

Capítulo 3

Modelo de rating

3.1 Introdução

A concessão de crédito traduz-se na disponibilização de um valor, geralmente por uma instituição financeira, mediante uma promessa de pagamento desse mesmo valor no futuro, que pressupõe a confiança na solvabilidade do devedor, isto é, de que o mesmo irá honrar os seus compromissos nas datas acordadas previamente. O risco de crédito é o risco de perda em que se incorre quando há incapacidade de pagamento numa operação de concessão de crédito.

Para avaliar este risco, e tentar discriminar bom de maus clientes, é elaborado um modelo de rating que conjuga duas avaliações: a qualitativa e a quantitativa. Esta consiste na atribuição de uma pontuação gerada por um modelo, por exemplo de regressão logística, de forma a determinar o risco de crédito de uma entidade. Este indicador é uma medida estatística que reflete a probabilidade da capacidade e da vontade de uma entidade honrar, atempadamente e na íntegra, os compromissos financeiros. O rating é calculado com base neste princípio e nas características das entidades registadas numa base de dados, refletindo a relação que estas características têm com a probabilidade de incumprimento (PD - default).

Em 1941, David Durand (ver [6]) realizou um estudo que se assemelha a um rating. Este estudo foi realizado para National Bureau of Economic Research (EUA), cujo objetivo era discriminar bons e maus empréstimos. Assim, o estudo analisou cerca de 7200 observações de empréstimos feitos a 37 empresas com base nas seguintes informações: idade, género, antiguidade no emprego, antiguidade na habitação, profissão, setor de atividade, contas bancárias, seguros de vida e valor da prestação mensal. Usando um teste qui-quadrado, conseguiu distinguir quais as variáveis significativas para distinguir bons de maus empréstimos. Durand usou uma função discriminante para criar um modelo de *credit scoring*.

Através de um modelo de regressão logística, determinado com base nas mais avançadas técnicas estatísticas, o rating atribuído a cada entidade reflete a interação de diversas variáveis financeiras, demográficas, de pagamento, de incidentes, entre outras.

Os modelos de regressão constituem uma das ferramentas estatísticas mais importantes na

análise de dados quando se pretende modelar relações entre variáveis. Esta técnica permite inferir a relação de uma variável dependente e um conjunto de variáveis, que se pretendem explicativas. A análise da regressão pode ser usada como um método descritivo da análise de dados (por exemplo, o ajustamento de retas e curvas) sem serem necessárias quaisquer suposições acerca dos processos que permitiram gerar os dados.

Os modelos lineares generalizados foram formulados, por volta de 1972, por John Nelder e Robert Wedderburn como uma maneira de unificar vários modelos estatísticos, incluindo a regressão linear e a regressão logística. Isto permitiu-lhes desenvolver um algoritmo geral para a estimativa de máxima verosimilhança em todos estes modelos.

Um dos casos particulares dos modelos lineares generalizados são os modelos onde a variável resposta apresenta apenas duas categorias ou que, de alguma forma, foi dicotomizada assumindo os valores 0 ou 1, sendo, possivelmente, o modelo de regressão logística o mais popular desses modelos. O principal objetivo destes é explorar a relação entre uma ou mais variáveis explicativas (ou independentes) e uma variável resposta (ou dependente).

A regressão logística é uma das técnicas estatísticas mais utilizadas no mercado para desenvolver modelos de classificação de empresas, vulgarmente conhecido por rating.

No decorrer deste capítulo, serão apresentadas algumas noções de regressão logística necessárias para a elaboração do modelo de rating pretendido. Apresentam-se também algumas técnicas para a seleção das variáveis do modelo de regressão logística que melhor se ajusta aos dados. Além disto, serão apresentadas métodos que permitem avaliar se o modelo obtido tem um bom ajuste.

3.2 O modelo de regressão logística

Nesta secção, serão apresentados conceitos da regressão logística que serão necessários para a perceção deste trabalho, com base em [9].

Os métodos de regressão são componente integrante de qualquer análise de dados onde se pretenda descrever a relação entre a variável dependente (ou resposta) e uma ou mais variáveis independentes (ou explicativas), com o objetivo de encontrar o melhor ajuste entre estas, usualmente chamadas covariáveis, e a variável resposta, permitindo explorar e inferir a relação entre elas. A análise de regressão pode ser usada como método descritivo da análise de dados sem serem necessárias quaisquer suposições acerca dos processos que permitiram gerar os dados.

O modelo de regressão mais comum é o linear, em que a aproximação é feita através de uma reta.

Um modelo de regressão que utilize apenas uma variável independente ou explicativa diz-se univariado; se o modelo utiliza mais do que uma variável explicativa diz-se multivariado.

3.2.1 Caso univariado

Considere-se duas variáveis X , e Y ; considerem-se ainda n pares ordenados $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ de valores observados destas variáveis. Se existir uma relação linear entre elas, então o modelo de regressão linear univariado que representa Y em função de X é:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \dots, n. \quad (3.1)$$

Neste modelo, Y_i é uma variável aleatória e representa o valor da variável resposta; x_i representa o valor da i -ésima observação da variável explicativa; ϵ_i é uma variável aleatória e representa o erro na i -ésima observação; e β_0 e β_1 são os parâmetros do modelo. Estes parâmetros são os valores se quer estimar num modelo de regressão.

Em qualquer problema de regressão, a quantidade fundamental é a média dos valores da variável resposta, dado o valor da variável independente, que se designa por média condicional e representa-se por $E(Y|x)$, sendo Y a variável resposta e x o valor da variável independente. Na regressão linear, a média condicional tem uma relação linear com os valores de X , sendo essa relação dada por:

$$E(Y|x) = \beta_0 + \beta_1 x. \quad (3.2)$$

A expressão referida indica que a média condicional nesta regressão pode tomar qualquer valor, isto é, $-\infty < E(Y|x) < +\infty$, dependendo da variação de x .

Considere-se agora que a variável resposta é dicotómica, isto é, Y toma apenas os valores 0 ou 1, que representam a ausência ou a presença de uma determinada característica.

A regressão logística é a técnica mais utilizada no mercado para o desenvolvimento de modelos de rating. Nas análises do Banco de Portugal é possível verificar que o risco de incumprimento - probabilidade de uma empresa não conseguir cumprir com as suas obrigações - é baixo para empresas com capacidade financeira forte, sendo que vai piorando para as classes de risco mais elevadas, o que leva a supor que a curva da função logística é uma boa aproximação para discriminar as empresas com maior ou menor probabilidade de incumprimento [15]. Além disto, na regressão logística não exige a suposição da normalidade das variáveis independentes e é mais robusta quando esta característica não é atendida.

No método da regressão logística, pretende-se encontrar um modelo adequado e parcimonioso que permita descrever a relação entre uma variável aleatória binária (sucesso ou insucesso) e um conjunto de variáveis não-aleatórias. O objetivo não é o de obter um modelo que discrimine totalmente os dois resultados possíveis (sucesso ou insucesso), mas sim que indique com precisão a probabilidade de ocorrência de um sucesso (ou de um insucesso). Nestes modelos, a variável dependente pode ser ordinal ou nominal e as independentes podem ser categóricas (desde que dicotomizadas após transformação) ou contínuas. Estas variáveis podem ser variáveis preditivas com características muito diferentes, podendo representar efeitos de curvatura e/ou interação. Esta flexibilidade torna o modelo de regressão logística multivariado muito atrativo.

A relação entre a variável independente e a variável dependente assemelha-se a:

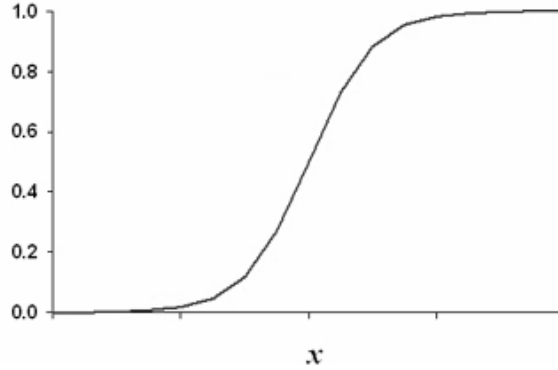


Figure 3.1: Gráfico da função logística

O modelo de regressão logística univariado é semelhante ao modelo de regressão linear univariado, no entanto, a sua variável resposta Y_i é binária. Deste modo, tem-se que

$$Y_i \sim B(p_i, n_i), \text{ para } i = 1, \dots, m \quad (3.3)$$

onde o número de ensaios de Bernoulli n_i é conhecido e a probabilidade de ocorrência p_i é desconhecida.

Considere-se Y_i uma variável aleatória de Bernoulli, ou seja, Y_i é uma variável binária, com distribuição de probabilidade tal que:

$$\begin{aligned} Y_i = 1 &\Rightarrow P(Y_i = 1) = \pi_i \\ Y_i = 0 &\Rightarrow P(Y_i = 0) = 1 - \pi_i. \end{aligned} \quad (3.4)$$

Usando a definição de valor esperado, tem-se que:

$$E(Y_i) = \pi_i. \quad (3.5)$$

Na regressão logística univariada, no caso em que a variável resposta toma 2 valores distintos, assume-se que $E(Y|x)$ pode ser escrito da seguinte forma:

$$\begin{aligned} E(Y|x) &= \pi(x) \\ &= \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \end{aligned} \quad (3.6)$$

e a razão dada por $\frac{\pi(x)}{1 - \pi(x)}$ é chamada de *odds*.

Uma transformação fulcral de $\pi(x)$ no estudo dos modelos de regressão logística é a *logit*, cujo objetivo é linearizar o modelo, aplicando o logaritmo. Essa transformação é definida pela seguinte expressão:

$$g(x) = \ln \left(\frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x \quad (3.7)$$

No modelo de regressão logística, pode-se expressar o valor da variável resposta dado o valor de x por $y = \pi(x) + \epsilon$, em que ϵ pode tomar um de dois valores: se $y = 1$, então $\epsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$; se $y = 0$, então $\epsilon = -\pi(x)$, com probabilidade $1 - \pi(x)$. Assim, neste caso, ϵ tem uma distribuição com média zero e variância $\pi(x)[1 - \pi(x)]$. Conclui-se então que a variável resposta condicionada segue uma distribuição binomial com probabilidade igual a $\pi(x)$.

Quando se está a construir um modelo de regressão que se ajuste a um certo conjunto de dados, é necessário estimar os valores dos parâmetros do modelo, que são desconhecidos. Assim, considere-se uma amostra de n observações independentes do par ordenado (x_i, y_i) , $i = 1, 2, \dots, n$, onde y_i representa o valor de uma variável resposta dicotômica e x_i é o valor da variável independente para a i -ésima observação. Assume-se que a variável resposta toma os valores 0 ou 1, representando a ausência ou a presença, respetivamente, de uma determinada característica. Para ajustar o modelo de regressão logística da equação 3.6 a um conjunto de dados, é necessário estimar os valores de β_0 e de β_1 , que são valores desconhecidos.

No modelo de regressão linear, o método utilizado para estimar os parâmetros é o dos mínimos quadrados, segundo o qual são escolhidos os valores de β_0 e de β_1 que minimizam a soma dos quadrados dos desvios entre os valores observados de Y e os valores obtidos pelo modelo. Os estimadores obtidos apresentam algumas características estatísticas desejáveis. No entanto, quando se aplica este método ao modelo de regressão logística, os estimadores não apresentam essas características.

O método utilizado para estimar os parâmetros β_0 e β_1 do modelo de regressão logística é o de máxima verosimilhança, método este que permite estimar os parâmetros de um modelo estatístico. Este método fornece estimadores para os parâmetros em estudo que maximizam a probabilidade de obter o conjunto de dados observados. Para aplicar este método, é necessário construir uma função, a função de verosimilhança, que exprime a probabilidade dos valores observados em função dos parâmetros desconhecidos. Os estimadores de máxima verosimilhança de β_0 e de β_1 são escolhidos para serem os valores que maximizam esta função. Assim, os estimadores resultantes são aqueles que se aproximam mais dos dados observados.

Para um valor arbitrário de $\boldsymbol{\beta} = (\beta_0, \beta_1)$, o vetor de parâmetros do modelo, se Y é um variável que toma os valores 0 ou 1, então a expressão para $\pi(x)$ dado na equação 3.6 fornece a probabilidade condicionada de Y ser igual a 1 dado x . Ou seja, $\pi(x) = P(Y = 1|x)$. Tem-se então que $1 - \pi(x)$ representa o valor da probabilidade condicionada de Y ser igual a 0 dado o valor de x , isto é, $1 - \pi(x) = P(Y = 0|x)$. Assim, para os pares ordenados (x_i, y_i) , com $y_i = 1$, a contribuição para a função de verosimilhança é $\pi(x_i)$ e, para os pares ordenados (x_i, y_i) , com $y_i = 0$, a contribuição para a função de verosimilhança é $1 - \pi(x_i)$, onde $\pi(x_i)$ representa o valor da função $\pi(x)$ para x_i . Uma forma conveniente de expressar a contribuição para a função de verosimilhança para o par ordenado (x_i, y_i) é

$$\pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}. \quad (3.8)$$

Uma vez que se assume que as observações são independentes, a função de verosimilhança é obtida como o produto dos termos dados na equação 3.8 do seguinte modo:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}. \quad (3.9)$$

O princípio de máxima verosimilhança estabelece que se utilize o estimador de $\boldsymbol{\beta}$ que maximiza a expressão 3.9. No entanto, matematicamente, é mais fácil trabalhar com o logaritmo desta equação. A expressão da função log-verosimilhança é definido por:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \quad (3.10)$$

Para encontrar o valor de $\boldsymbol{\beta}$ que maximiza $L(\boldsymbol{\beta})$, tem de se derivar esta função em relação a β_0 e β_1 e igualar cada uma das expressões obtidas a zero. As equações obtidas, conhecidas como equações de verosimilhança, são:

$$\sum_{i=1}^n [y_i - \pi(x_i)] = 0 \quad (3.11)$$

e

$$\sum_{i=1}^n x_i [y_i - \pi(x_i)] = 0 \quad (3.12)$$

Estas equações não são lineares em relação a β_0 e β_1 , pelo que são utilizado métodos iterativos para resolvê-las. Deste modo, estas equações são resolvidas utilizando eficientemente os recursos informáticos. O valor de $\boldsymbol{\beta}$ que é solução das equações 3.11 e 3.12 é chamado de estimador de máxima verosimilhança.

3.2.2 Caso multivariado

O modelo de regressão logística multivariado é semelhante ao modelo de regressão logística univariado, no entanto a variável resposta depende de várias variáveis explicativas.

Considere-se um conjunto de p variáveis independentes denotadas pelo vetor $\mathbf{x}' = (x_1, x_2, \dots, x_p)$. Seja a probabilidade condicionada de Y ser 1 denotada por $P(Y = 1|x) = \pi(x)$. A transformação logit do modelo de regressão logística multivariado é dado pela equação:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p, \quad (3.13)$$

em que os coeficientes $\beta_0, \beta_1, \dots, \beta_p$ são os parâmetros do modelo; para este caso, o modelo de regressão logística é dado por:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}. \quad (3.14)$$

ou seja, $g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$.

Analogamente ao caso univariado, é necessário estimar os valores dos parâmetros do modelo. Assume-se que se tem um conjunto de n observações independentes (\mathbf{x}_i, y_i) , com $i = 1, 2, \dots, n$, sendo $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ o vetor da i -ésima observação do vetor \mathbf{x}' . Pretende-se agora estimar o vetor $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_p)$. O método utilizado para a estimação é, em semelhança ao caso univariado, o de máxima verosimilhança. A função de verosimilhança é semelhante à dada na equação 3.9, no entanto $\pi(x)$ é definido como na equação 3.14. Diferenciando a função log-verosimilhança em relação aos $p + 1$ parâmetros do modelo, obtêm-se $p + 1$ equações de verosimilhança que se representam do seguinte modo:

$$\sum_{i=1}^n [y_i - \pi(\mathbf{x}_i)] = 0 \quad (3.15)$$

e

$$\sum_{i=1}^n x_{ij} [y_i - \pi(\mathbf{x}_i)] = 0 \quad (3.16)$$

com $j = 1, 2, \dots, p$. Tal como foi referido no caso univariado, a resolução destas equações de verosimilhança requerem a utilização eficiente de recursos informáticos.

3.3 Análise preditiva do modelo

Ao seleccionarmos modelos, é preciso ter em mente que não existem modelos verdadeiros. Há apenas modelos aproximados da realidade que causam perda de informações. Deste modo, é necessário fazer a escolha do “melhor” modelo, de entre aqueles que foram ajustados, para explicar o fenómeno sob estudo.

Segundo [17], a necessidade da obtenção de medidas de discriminação precisas é um problema que remonta ao século XIX. Gustav Theodor Fechner foi considerado o pioneiro, com desenvolvimentos no campo da psicologia sensorial. Na área da estatística este problema em termos de testes de hipóteses, ou tomada de decisões, foram desenvolvidos por vários autores, Thurstone, Blackwell, e posteriormente Neyman e Pearson.

Apresentam-se a seguir três critérios que ajudam a procurar o modelo que melhor se ajusta a um conjunto de dados em estudo.

3.3.1 Teste de Wald

Segundo [9], em regressão logística, tem-se variáveis resultado e uma ou mais variáveis explicativas. Para cada variável explicativa do modelo, haverá um parâmetro associado. O

teste de Wald é usado para aferir a significância de cada um dos coeficientes do modelo.

Considere-se o modelo de regressão logística univariado tal que a função logit é dada por $g(x) = \beta_0 + \beta_1 x$; quer-se testar a significância da variável x para a variável resposta. O teste de Wald é obtido através da razão entre estimador de máxima verosimilhança do parâmetro β_1 , $\hat{\beta}_1$, e o estimador do seu desvio-padrão. A razão resultante, sob a hipótese $\beta_1 = 0$, segue uma distribuição normal padrão, $N(0, 1)$. Ou seja, a estatística do teste de Wald é dada pela expressão:

$$W = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}$$

e o p -valor é definido através da expressão:

$$p\text{-valor} = P(|Z| > W),$$

em que Z representa uma variável aleatória tal que $Z \sim N(0, 1)$.

O caso do modelo multivariado é semelhante ao caso univariado. Assim, o teste de Wald para o parâmetro β_j é obtido através da razão entre o estimador de máxima verosimilhança do parâmetro ($\hat{\beta}_j$) e a estimativa do seu desvio-padrão. Isto é, considerando as seguintes hipóteses

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases},$$

a estatística de teste é

$$W_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}.$$

Considerando um certo nível de significância, se a hipótese H_0 for aceite, tem-se que a variável X_j não explica a variável resposta.

Este quociente, sob a hipótese nula, segue uma distribuição normal padrão ($N(0, 1)$). O p -valor é definido como

$$p\text{-valor} = P[|Z| \geq |W_j| | H_0] = 2P[Z \geq |W_j| | H_0],$$

tendo que $Z \sim N(0, 1)$.

A performance deste teste foi avaliada por vários autores que concluíram que apresenta algumas limitações:

- requer o cálculo do estimador de máxima verosimilhança de β_1 ;
- muitas vezes aceita a hipótese nula, quando o coeficiente é significativo.

Em R , o p -valor do teste de Wald é obtido após correr o comando *glm*.

3.3.2 Critério de Informação de Akaike

Segundo [1], em 1974, Hirotugu Akaike apresentou um critério que ajuda a selecionar, de entre vários modelos, o melhor, tendo apenas em conta a função de máxima verosimilhança e o número de variáveis presentes no modelo. Assim, era possível desprender esta decisão das ambiguidades inerentes à utilização de testes de hipóteses.

Akaike utilizou a Informação de Kullback-Leibler para testar se um dado modelo é adequado. Porém o seu uso é limitado, uma vez que depende da distribuição do modelo verdadeiro, que é na prática desconhecida.

O Critério de Informação de Akaike (AIC) é definido como:

$$AIC_n = -2 * (\log(l) - k) \quad (3.17)$$

em que l é a função de máxima verosimilhança do modelo e k é o número de parâmetros independentes ajustados no modelo.

Assim, objetivo é balancear o ajuste do modelo com a quantidade de variáveis, que é tanto melhor quanto menor é o valor do AIC.

Em R, o valor do AIC é obtido através do comando *aic*.

Seleção de variáveis

Segundo [9], a abordagem tradicional na construção de modelos estatísticos é encontrar o modelo mais parcimonioso que explica os dados. Quanto mais variáveis no modelo, maior se torna a estimativa do erro e mais dependente o modelo fica dos dados observados.

O método Stepwise para a seleção de variáveis é muito usado na seleção de variáveis para modelos de regressão.

Qualquer procedimento para seleção ou exclusão de variáveis de um modelo é baseado em um algoritmo que verifica a importância das variáveis, incluindo ou excluindo-as do modelo baseando-se numa regra de decisão. A importância da variável é definida em termos de uma medida de significância estatística do coeficiente associado à variável para o modelo. Esse procedimento parte da suposição inicial de que todas as variáveis fazem parte do modelo, elimina-se ou adiciona-se uma variável de cada vez, utilizando o valor do AIC, ou seja, quanto menor é o valor deste critério melhor é o modelo.

Em R, o método stepwise para escolha de um modelo com base no AIC é obtido através do comando *step* em que *direction="both"*.

3.3.3 Curva de ROC

A análise da curva ROC (Receiver Operating Characteristic) teve origem na teoria de decisão estatística e foi desenvolvida entre 1950 e 1960 para avaliar a detecção de sinais em radares e na

psicologia sensorial.

A área abaixo da curva ROC dá uma descrição mais completa da precisão da classificação dada pelo modelo. O gráfico desta curva relaciona a probabilidade de ser detectado um sinal verdadeiro, que é obtida através do cálculo da "sensibilidade" do modelo, e de um sinal falso, que é obtida através do cálculo de "1-especificidade", para um conjunto de pontos de corte (*cutoff point*).

A sensibilidade de um modelo é a proporção de verdadeiros positivos, ou seja, traduz a capacidade do sistema em prever corretamente a presença da condição em análise para casos que realmente a apresentam. Tem-se então que:

$$\text{Sensibilidade} = \frac{\text{Número de acertos positivos}}{\text{Número total de positivos}}$$

A especificidade de um modelo é a proporção de verdadeiros negativos, ou seja, traduz a capacidade do sistema em prever corretamente a ausência da condição em análise para casos que realmente não a apresentam. Tem-se então que:

$$\text{Especificidade} = \frac{\text{Número de acertos negativos}}{\text{Número total de negativos}}$$

No caso de uma resposta binária, o ponto de corte ou *cutoff point* é o valor para os quais os valores acima são classificados como evento positivo e os valores abaixo como evento negativo.

A precisão do modelo é a proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. É dado por:

$$\text{Precisão} = \frac{\text{Número de acertos negativos} + \text{Número de acertos positivos}}{\text{Número total de observações}}$$

Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema.

As curvas ROC permitem quantificar a exatidão de um teste diagnóstico, já que, esta é proporcional à área sob a curva ROC, isto é, tanto maior quanto mais a curva se aproxima do canto superior esquerdo do diagrama. Sabendo isto, a curva será útil, também, na comparação de testes diagnósticos, tendo um teste uma exatidão tanto maior, quanto maior for a área sob a curva ROC.

Na figura seguinte, apresentam-se três exemplos de curvas de ROC para diferentes modelos: A - excelente poder discriminatório, B - aceitável poder discriminatório, e C - sem poder discriminatório.

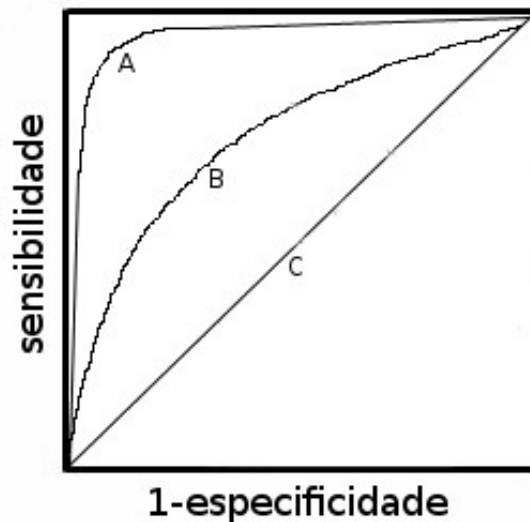


Figure 3.2: Curva de ROC

Como regra geral, tem-se (ver [9]):

$ROC = 0.5$	Sem poder discriminatório
$0.5 < ROC < 0.7$	Baixo poder discriminatório
$0.7 \leq ROC < 0.8$	Aceitável poder discriminatório
$0.8 \leq ROC < 0.9$	Bom poder discriminatório
$0.9 \leq ROC$	Excelente poder discriminatório

Em R, é possível obter o gráfico desta curva, através do comando *roc*, assim como a área abaixo da curva através do comando *auc*.

3.4 Amostra de treino e amostra de teste

Para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados, é usada a técnica validação cruzada (ver [12]). Esta técnica é utilizada em problemas onde o objetivo do modelo é a predição. Procura-se então estimar o quão preciso é o modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

Um dos métodos consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treino, ou seja, para estimar dos parâmetros do modelo, e outro para teste, para validação do modelo. O conjunto de dados pode ser separado em quantidades iguais ou não. Uma proporção muito comum é considerar dois terços dos dados para treino e o restante para teste. Neste trabalho, optou-se por colocar 70% dos dados para treino e 30% para teste.

As duas amostras criadas por este método têm de ser representativas da população do

problema em questão. Assim, após se definir as amostras, há que verificar se a proporção dos dados entre classes se mantém.

Considere-se duas variáveis aleatórias X e Y que representam uma determinada característica de duas populações com distribuição Bernoulli com parâmetros p_1 e p_2 , respetivamente. Seleccionam-se duas amostras independentes, X_1, X_2, \dots, X_{n_1} e Y_1, Y_2, \dots, Y_{n_2} , dessas populações. Sabe-se que cada X_i , com $i = 1, \dots, n_1$, e cada Y_j , com $j = 1, \dots, n_2$, tem distribuição Bernoulli com parâmetros p_1 e p_2 , respetivamente, ou seja:

$$X_1, \dots, X_{n_1} \sim \text{Bernoulli}(p_1) \text{ e } Y_1, \dots, Y_{n_2} \sim \text{Bernoulli}(p_2)$$

com médias p_1 e p_2 e variâncias $\sigma_1^2 = p_1(1 - p_1)$ e $\sigma_2^2 = p_2(1 - p_2)$, respetivamente.

As variáveis $\hat{p}_1 = \bar{X}$ e $\hat{p}_2 = \bar{Y}$ são estimadores de máxima verosimilhança para p_1 e p_2 , se cada categoria tiver mais de 5 elementos, respetivamente, e têm distribuição amostral aproximadamente normal, ou seja:

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{e} \quad \hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

Tem-se então que:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

isto é,

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1).$$

Para se realizar o teste para duas proporções com aproximação Normal, vai-se considerar a hipótese nula $p_1 = p_2$. Assim, sob a hipótese nula, $\hat{p}_1 - \hat{p}_2$ tem distribuição Normal com média $\mu = 0$ e desvio padrão

$$\sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}$$

onde $p = p_1 = p_2$.

Como não se conhece o valor p , utilizar-se uma média ponderada de \hat{p}_1 e \hat{p}_2 para se obter um valor estimado daquele valor:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Assim, o valor de σ é obtido da seguinte forma:

$$\sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

Portanto, tem-se que

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \sim N(0, 1).$$

Posto isto, considerando o teste de hipóteses $H_0 : p_1 = p_2$ contra $H_1 : p_1 \neq p_2$ e fixado um nível de significância α , rejeita-se H_0 se

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} > \phi\left(\frac{\alpha}{2}\right)$$

ou

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} < -\phi\left(\frac{\alpha}{2}\right)$$

em que ϕ é a função distribuição da normal padrão, isto é, $N(0, 1)$.

3.5 Escalões de rating

Para criar os escalões de rating, após alguma pesquisa online, descobriu-se algumas respostas para as quais se necessitava de resposta.

- Quantos escalões são precisos?
- Pretende-se uma resposta intermédia, ou seja, um escalão que fique exatamente a meio da escala?
- Como se deve classificar o escalão de forma a ser fácil interpretar o seu significado?
- Como construir estes escalões?

Na pesquisa efetuada, foram encontradas várias respostas para estas perguntas, pelo que não existe uma única forma de se estabelecer os escalões para o rating. Assim, as respostas àquelas perguntas ficam ao critério do problema em estudo.

Quanto ao número de escalões e ao facto de se ter ou não um escalão intermédio, optou-se por 7 níveis de rating, assim, além de ser possível estabelecer o referido escalão intermédio - escalão 4, Razoável -, permite a existência de diferenciação entre as empresas sem demasiados níveis. Caso se optasse por ter mais níveis de rating, devido à forma de construção do modelo que "tenta" colocar as piores empresas, isto é, com maior probabilidade de incumprimento (PD), nos últimos escalões, o rating poderia torna-se pouco fiável, uma vez que, para PD's mais baixas, o incumprimento verificado pode ser o mesmo quando estes escalões ficam com populações demasiado pequenas.

Quanto à denominação de cada escalão, a seleção também não é única. As agências de rating classificam as empresas com diferentes número de escalões e denominações, conforme pode ser verificado no anexo da instrução número 16/2013 do Banco de Portugal (BdP) (disponível em [4]):

BdP	Fitch	Moody's	S&P	Coface
1	AAA até AA-	Aaa até Aa3	AAA até AA-	10 até 9
2	A+ até A-	A1 até A3	A+ até A-	8
3	BBB+ até BBB-	Baa1 até Baa3	BBB+ até BBB-	7 até 6
4	BB+ até BB-	Ba1 até Ba3	BB+ até BB-	5 até 4
5	B+ até B-	B1 até B3	B+ até B-	3
6	Inferior a B-	Inferior a B3	Inferior a B-	2 até 1

Table 3.1: Escalões de rating reconhecidos pelo Banco de Portugal para avaliação externa

BdP	ICAP	CPR	Informa D&B	DBRS
1	—	—	12	AAA até AAL
2	AA, A	AAA até AA-	11	AH até AL
3	BB, B	A+ até A-	10 até 8	BBBH até BBBL
4	C, D, E	BBB+ até BBB-	7 até 6	BBH até BBL
5	F	BB+ até BB-	5 até 4	BH até BL
6	G, H	Inferior a BB-	3 até 1	Inferior a BL

Table 3.2: Escalões de rating reconhecidos pelo Banco de Portugal para avaliação externa

Neste caso optou-se por:

Rating
1 - Formidável
2 - Excelente
3 - Bom
4 - Razoável
5 - Inadequado
6 - Fraco
7 - Terrível

Table 3.3: Escalões de rating utilizados

Estas denominações dos escalões de rating permitem criar uma boa ordenação das empresas e facilita a comunicação, entre pessoas, sobre a qualidade de um cliente/empresa.

Para a construção dos escalões, decidiu-se dividir o número de clientes por escalão de forma homogénea. Existem outras formas de fazer esta divisão: a título de exemplo, pode-se dividir os clientes de forma a que o número de clientes por escalão seguisse uma distribuição normal ou utilizar a probabilidade de incumprimento esperada de cada cliente obtida pelo modelo.

3.6 Variáveis qualitativas

Existem outras variáveis, de índole qualitativa, que podem analisar a qualidade de uma empresa e que podem não estar refletidas nas contas da mesma.

Informações como experiência e influência do empresário no setor, concorrência no setor, expectativas, confiança transmitida e capacidade de execução dos projetos não existem nas contas. Por este facto, pode-se criar um conjunto de questões com resposta por escalões para apurar o rating mais ajustado para uma determinada empresa.

Não foi possível, devido à confidencialidade dos dados, requisitar aos gestores da instituição financeira a opinião sobre um subconjunto das empresas, de forma a definir um modelo qualitativo. Propõe-se que, para a implementação deste modelo na instituição financeira, sejam elaboradas as questões relacionadas com experiência e influência do empresário no setor, concorrência no setor, expectativas, confiança transmitida e a capacidade de execução dos projetos, de forma a obter outras informações sobre o risco da empresa que é explicado pelas contas e que será do conhecimento dos gestores. Estas variáveis qualitativas, posteriormente, poderão ser utilizadas para a construção do modelo junto com as variáveis contínuas e fazer um ajuste ao mesmo.

Tal como sugerido, estas variáveis podem ser no futuro incluídas no modelo, através da seleção de variáveis da regressão pelo método AIC ou como um modelo alternativo com uma ponderação limitada. Denotando por p a ponderação, por r a pontuação obtida através do modelo da regressão e por q a pontuação obtida através do modelo qualitativo; o escalão de rating desta sugestão seria dado por:

$$\text{Pontuação Final} = r * (1 - p) + q * (p)$$

Capítulo 4

Aplicação do modelo de rating

4.1 Algoritmo aplicado para selecção do modelo de rating

Neste trabalho, para elaboração do modelo que melhor se aproxima aos dados reais utilizou-se o programa R., criado originalmente por Ross Ihaka e por Robert Gentleman. O R, com a utilização do interface do RStudio, é uma ferramenta que pode ser utilizada para diversos cálculos estatísticos e análises gráficas.

Foram, ainda, utilizados alguns pacotes do R que serão mencionados, ao longo dos algoritmos constantes neste trabalho.

i. Preparação dos dados

No processo de preparação dos dados procurou-se identificar padrões entre os mesmos. Para tal aplicaram-se algumas técnicas de *data mining* para fornecer o melhor conjunto de dados para as previsões de incumprimento das empresas.

(a) Criar uma amostra de teste e uma amostra de treino

Separar o universo de clientes em duas amostras, uma de treino e outra de teste. A amostra de treino irá servir para utilizar na regressão logística e obter um modelo; posteriormente, a amostra de teste servirá para validar a qualidade do modelo final.

(b) Validação das amostras

Verificar que as amostras mantêm a mesma proporção de empresas em incumprimento por CAE e por categoria.

(c) Correr a regressão logística univariada

Ao aplicar a regressão logística univariada pretende-se perceber a capacidade explicativa de cada uma das variáveis isoladamente.

(d) Análise das correlações

São retiradas todas as variáveis com uma correlação muito forte e forte (positiva ou negativa).

- ii. Utilização do método stepwise para seleção de variáveis para a regressão logística multivariada
 - (a) Utilização do método de eliminação backward para seleção de variáveis para a regressão logística multivariada
 - (b) Utilização do método de selecção forward para seleção de variáveis para a regressão logística multivariada
 - (c) Utilização dos métodos forward e backward (comando do R: both) na para seleção de variáveis para a regressão logística multivariada
- iii. Comparativo das variáveis seleccionadas em cada método com os valores obtidos para a regressão logística univariada.

Permite averiguar se as variáveis finais escolhidas pelos métodos forward, backward e both são as variáveis com a melhor capacidade explicativa quando se aplica a regressão logística univariada.

 - (a) Análise do Teste de Wald
Verificação dos valores obtidos no teste de Wald
 - (b) Análise da curva de ROC
Análise dos valores de sensibilidade e especificidade do modelo através da curva de ROC.
 - (c) Análise comparativa dos métodos
Comparação dos testes para cada um dos métodos utilizados.
- iv. Divisão em escalões de rating
- v. Análise da taxa de execução por escalão de rating

4.2 Aplicação do modelo de regressão logística aos dados

Nesta análise foi mantido o respeito à confidencialidade dos dados utilizados, não sendo, por isso, possível saber o nome das variáveis utilizadas para a regressão logística.

Utilizou-se o CAE e a categoria da empresa para verificar se a amostra de treino e a amostra de teste mantinham características idênticas ao nível das proporções de empresas com cumprimento e em incumprimento. Para isso, recorreu-se ao teste da binomial, com um nível de significância de 0.05.

Categoria	Treino		Teste		Teste Binomial
	Em Cumpr.	Incumpr.	Em Cumpr.	Incumpr.	Z-value
Grande	478	18	120	6	5.289
Média	4892	228	2104	101	-0.301
Pequena	24421	940	10288	449	1.006
Micro	50259	1632	21627	729	-1.600

Table 4.1: Distribuição das empresas pelas amostras de treino e teste, por categoria

O teste da binomial, não rejeita a hipótese de as proporções na amostra de treino serem iguais à amostra de teste, com exceção para as grandes empresas. Devido ao número pouco significativo de empresas, a mesma não foi tomada em consideração.

CAE	Treino		Teste		Teste Binomial
	Em Cumpr.	Incumpr.	Em Cumpr.	Incumpr.	Z-value
A	1246	19	502	11	1.063
B	280	16	135	6	-1.036
C	15449	512	6558	6 207	0.848
D,E	365	7	161	5	-0.434
F	7650	456	3247	216	0.163
G	28411	985	12258	457	-1.085
H	3686	137	1566	55	0.369
I	4942	187	2101	80	0.315
J	1561	54	690	20	-0.572
K	371	12	155	3	0.404
L	1201	46	545	22	-1.178
M,N	9111	263	3889	133	-0.065
P,Q	4198	72	1787	39	0.079
R,S,U	1579	52	635	31	1.061

Table 4.2: Distribuição das empresas pelas amostras de treino e teste, por CAE

Neste caso, o teste da binomial, não rejeita a hipótese de as proporções na amostra de treino serem iguais à amostra de teste.

Depois de definir a amostra de treino e a amostra de teste e fazer a validação das proporções de empresas em cumprimento e em incumprimento, estabeleceu-se um modelo de regressão logística univariado para cada uma das variáveis. Destes modelos, analisou-se o valor do AIC em cada um; estes valores são apresentados na tabela seguinte:

Variável	AIC	Variável	AIC	Variável	AIC
VAR29	23271.22	VAR09	24533.56	VAR25	24592.77
VAR40	23629.75	VAR10	24533.56	VAR70	24593.19
VAR60	23740.79	VAR45	24533.56	VAR63	24593.9
VAR32	24050.73	VAR62	24537.37	VAR11	24594.55
VAR31	24094.38	VAR55	24539.74	VAR15	24594.65
VAR39	24104.43	VAR54	24544.7	VAR24	24595.13
VAR66	24109.58	VAR08	24545.64	VAR17	24596.88
VAR38	24119.56	VAR44	24553.61	VAR23	24597.71
VAR46	24201.73	VAR30	24557.54	VAR01	24599.29
VAR37	24321.97	VAR04	24567.59	VAR27	24599.52
VAR28	24331.18	CATEG	24570.18	VAR02	24599.61
VAR36	24337.21	VAR13	24572.18		
VAR51	24371.29	VAR03	24573.44		
VAR49	24389.6	VAR07	24573.92		
VAR67	24433.3	VAR53	24578.57		
CAE	24436.03	VAR19	24585.03		
VAR58	24439.14	VAR20	24586.52		
VAR43	24461.32	VAR05	24586.7		
VAR59	24474.09	VAR21	24587.68		
VAR64	24476.29	VAR47	24587.79		
VAR52	24508.11	VAR33	24588.7		
VAR57	24531.06	VAR34	24590.57		
VAR68	24533.16	VAR48	24591.07		

Table 4.3: AIC do modelo de regressão logística univariado

Apesar do CAE estar na 16 posição no nível individual, os dados recebidos da instituição financeira são referentes a anos de crise em Portugal, onde o sector F teve imensas dificuldades e inúmeras empresas do sector abriram insolvência devido à conjuntura. Além disso, a instituição financeira que enviou os dados tem a sua carteira repartida regionalmente e o rating será para servir para todas as regiões e não deverá discriminar uma região em relação a outra, não existindo forma de provar que os CAE estão repartidos de forma equitativa pelo país, por exemplo a zona norte é uma zona industrial, enquanto a zona do Algarve é uma zona turística, por este facto decidiu-se excluir da regressão logística multivariada.

Por razões de cariz económico foram retiradas 17 variáveis com este processo, a saber: VAR06, VAR12, VAR14, VAR16, VAR18, VAR22, VAR26, VAR35, VAR41, VAR42, VAR50, VAR56, VAR61, VAR65, VAR69, VAR71 e VAR72.

Tendo em conta as restantes variáveis, e para se seleccionar aquelas que são mais significativas para a variável incumprimento, utilizou-se o método *both*, disponível na função *step* do R.

Foram testados diversos modelos; aqueles que se mostraram mais interessantes para a análise e com melhor comportamento foram os que consideram a separação em dois modelos distintos definidos pelo tipo de empresas: um para micro empresas e outro para pequenas, médias e grandes empresas.

4.2.1 Modelo construídos a partir da base de dados completa

Nesta primeira fase, foram utilizadas todas as restantes variáveis, independentemente da correlação existente entre elas. Tendo em conta os vários modelos obtidos, apresentam-se a seguir que foram considerados mais relevantes.

Para o modelo das micro empresas temos os seguintes coeficientes:

```

Coefficients:
              z value Pr(>|z|)
(Intercept) -17.922 < 2e-16 ***
VAR03        -3.463 0.000534 ***
VAR04         2.683 0.007297 **
VAR05        -2.506 0.012202 *
VAR07         2.777 0.005481 **
VAR13        -3.000 0.002703 **
VAR17         3.882 0.000104 ***
VAR19        -2.718 0.006570 **
VAR24         1.547 0.121785
VAR27        -3.074 0.002115 **
VAR29         3.694 0.000220 ***
VAR30        -6.342 2.27e-10 ***
VAR33         5.535 3.11e-08 ***
VAR34         2.785 0.005352 **
VAR39        -4.894 9.88e-07 ***
VAR40         6.682 2.36e-11 ***
VAR43         6.517 7.15e-11 ***
VAR44        -3.261 0.001108 **
VAR46        -4.317 1.58e-05 ***
VAR47        -2.530 0.011417 *
VAR48        -1.543 0.122841
VAR49        -6.248 4.16e-10 ***
VAR51         3.996 6.45e-05 ***
VAR52         3.802 0.000144 ***
VAR55        -4.690 2.73e-06 ***
VAR59        -1.744 0.081242 .
VAR63         5.214 1.85e-07 ***
VAR64         2.016 0.043820 *
VAR68        -3.049 0.002296 **
VAR70         3.365 0.000765 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 4.1: Coeficientes do modelo das micro empresas

A seguir apresenta-se a curva de ROC para este modelo. A área por baixo desta curva é de 0.744.

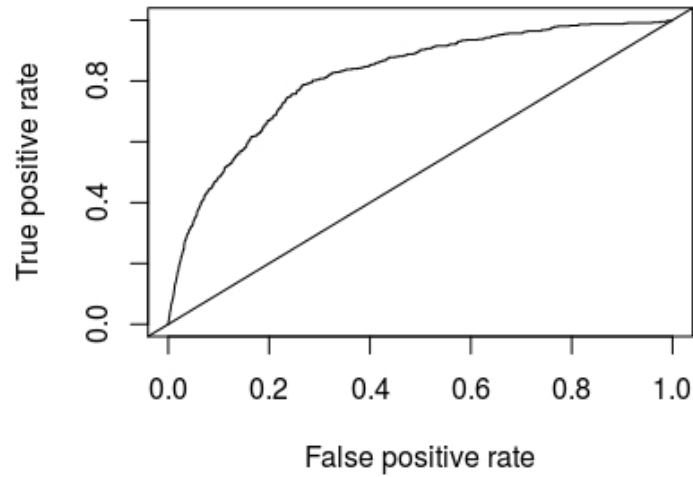


Figure 4.2: Análise da curva de ROC do modelo para as micro empresas

De seguida, apresenta-se o gráfico da medida de precisão. Esta medida otimiza os resultados corretos, contudo pode ser distorcida se a diferença entre o número de empresas cumpridoras e em incumprimento for muito elevada.

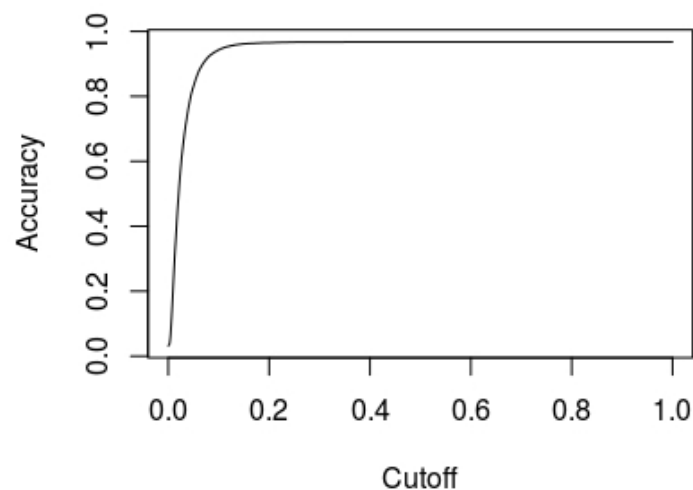


Figure 4.3: Análise da curva de ROC do modelo para as micro empresas

Conforme analisado no capítulo 2, o número de cumprimentos é muito superior ao número de incumprimentos, pelo que a medida de precisão pode estar distorcida.

Elaborou-se uma análise do comportamento do modelo removendo as variáveis com p-value menos significativo.

p-value	AIC	ROC
Original	13 226.70	0.7441
]0.1; 1]	13 228.76	0.7437
]0.05; 0.1]	13 230.15	0.7431

Table 4.4: AIC e ROC dos modelos com as variáveis sem restrições de correlações

Tendo por base o observado na tabela acima, decidiu-se optar pelo modelo original - sem remoção de variáveis -, uma vez que este apresenta os melhores valores do AIC e ROC, apesar das diferenças não serem significativas.

As taxas, esperadas e observadas, do modelo escolhido, de execução por escalão de *rating* foram as seguintes:

Rating	Esperado	Em cumprimento	Incumprimento	Observado	Diferença
1 - Formidável	0.0063	2962	24	0.0080	-0.0017
2 - Excelente	0.0101	3064	25	0.0081	0.0021
3 - Bom	0.0165	3119	55	0.0173	-0.0008
4 - Razoável	0.0189	3630	71	0.0191	-0.0003
5 - Inadequado	0.0279	2980	93	0.0302	-0.0023
6 - Fraco	0.0421	3068	144	0.0448	-0.0027
7 - Terrível	0.0999	2804	317	0.1015	-0.0016

Table 4.5: Escalões de rating para as micro empresas

Para o modelo das pequenas, médias e grandes empresas temos os seguintes coeficientes:

```

Coefficients:
              z value Pr(>|z|)
(Intercept) -18.446 < 2e-16 ***
VAR01       -1.521  0.12837
VAR05       -1.346  0.17828
VAR07        1.870  0.06151 .
VAR09       -2.357  0.01841 *
VAR13        4.888 1.02e-06 ***
VAR15        4.014 5.96e-05 ***
VAR19       -2.921  0.00349 **
VAR20       -3.177  0.00149 **
VAR21        1.893  0.05832 .
VAR24       -1.356  0.17518
VAR28       -2.503  0.01231 *
VAR31       -2.800  0.00511 **
VAR34        2.970  0.00298 **
VAR37       -1.432  0.15206
VAR39        3.000  0.00270 **
VAR40       10.422 < 2e-16 ***
VAR43        5.127 2.95e-07 ***
VAR46       -4.282 1.85e-05 ***
VAR47       -2.402  0.01632 *
VAR48       -2.104  0.03539 *
VAR51        6.081 1.20e-09 ***
VAR53        4.419 9.91e-06 ***
VAR54       -1.498  0.13402
VAR55        2.345  0.01903 *
VAR59       -5.835 5.39e-09 ***
VAR60        5.453 4.96e-08 ***
VAR63        4.193 2.75e-05 ***
VAR64       -5.942 2.81e-09 ***
VAR66        2.951  0.00317 **
VAR68       -1.970  0.04887 *
VAR70        2.269  0.02325 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 4.4: Coeficientes do modelo das pequenas, médias e grandes empresas

A área por baixo da curva de ROC obtida para o modelo das pequenas, médias e grandes empresas é de 0.826. A curva de ROC para este modelo apresenta-se a seguir.

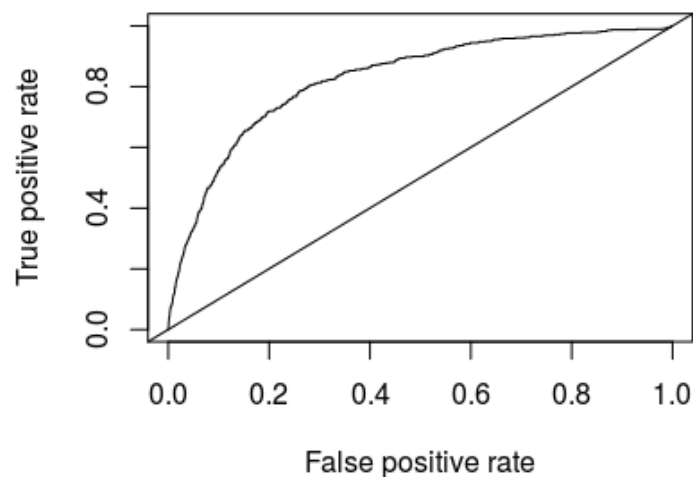


Figure 4.5: Análise da curva de ROC do modelo para as micro empresas

De seguida, apresenta-se o gráfico da medida de precisão para este modelo.

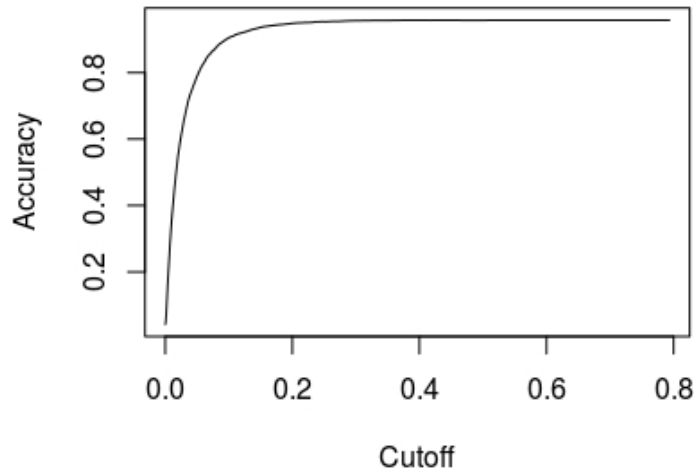


Figure 4.6: Análise da curva de ROC do modelo para as micro empresas

Conforme analisado no capítulo 2, o número de cumprimentos é muito superior ao número de incumprimentos. Apesar de analisado o gráfico da medida de precisão para todos os modelos, o mesmo não acrescenta nada de significativo para a análise.

Em semelhança ao que foi feito para as micro empresas elaborou-se a análise do comportamento do modelo removendo as variáveis com p-value menos significativo.

p-value	AIC	ROC
Original	8 490.011	0.8265
]0.1; 1]	8 493.93	0.8264
]0.05; 0.1]	8 493.33	0.8265

Table 4.6: AIC e ROC dos modelos com as variáveis sem restrições de correlações

Tendo por base o observado, decidiu-se optar pelo modelo original - sem remoção de variáveis -, uma vez que este apresenta os melhores valores do AIC e ROC, apesar das diferenças não serem significativas.

As taxas, esperadas e observadas, do modelo original, de execução por escalão de *rating* foram as seguintes:

Rating	Esperado	Em cumprimento	Incumprimento	Observado	Diferença
1 - Formidável	0.0046	1818	11	0.0060	-0.0014
2 - Excelente	0.0057	1826	10	0.0054	0.0003
3 - Bom	0.0110	1814	18	0.0098	0.0012
4 - Razoável	0.0177	2061	33	0.0157	0.0020
5 - Inadequado	0.0290	1804	47	0.0253	0.0037
6 - Fraco	0.0539	1739	112	0.0605	-0.0066
7 - Terrível	0.1487	1540	325	0.1742	-0.0255

Table 4.7: Escalões de rating para as pequenas, médias e grandes empresas

4.2.2 Modelos construídos sem correlações superiores a 0,90

Após a análise das correlações entre as variáveis, foram retiradas aquelas cuja correlação entre elas é superior a 0,9, sendo mantida apenas a variável com melhor AIC no modelo de regressão logística univariado. Desta análise, foram retiradas as 11 variáveis seguintes: VAR04, VAR05, VAR08, VAR10, VAR21, VAR24, VAR27, VAR31, VAR33, VAR38 e VAR45.

Tendo em conta as variáveis restantes, construíram-se novos modelos para micro empresas e para pequenas, médias e grandes empresas.

Para o modelo das micro empresas, tem-se os seguintes coeficientes:

```

Coefficients:
              zvalue Pr(>|z|)
(Intercept)  18.331 < 2e-16 ***
VAR02        -2.811 0.004942 **
VAR03        -2.763 0.005722 **
VAR07         2.174 0.029718 *
VAR09         1.908 0.056364 .
VAR13        -2.108 0.035073 *
VAR15        -1.531 0.125759
VAR17         3.368 0.000758 ***
VAR19        -2.393 0.016721 *
VAR29         4.142 3.45e-05 ***
VAR30        -4.567 4.94e-06 ***
VAR34         4.233 2.30e-05 ***
VAR36         2.650 0.008053 **
VAR37         2.509 0.012090 *
VAR39        -3.018 0.002546 **
VAR40         9.336 < 2e-16 ***
VAR43         6.156 7.45e-10 ***
VAR44        -4.177 2.95e-05 ***
VAR46        -4.243 2.21e-05 ***
VAR47        -2.683 0.007292 **
VAR48        -2.084 0.037172 *
VAR49        -6.072 1.26e-09 ***
VAR51         2.577 0.009966 **
VAR52         3.670 0.000243 ***
VAR55        -4.795 1.63e-06 ***
VAR59        -1.996 0.045977 *
VAR63         4.097 4.18e-05 ***
VAR68        -2.956 0.003112 **
VAR70         3.269 0.001078 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 4.7: Coeficientes do modelo das micro empresas

A seguir apresenta-se a curva de ROC para este modelo. A área por baixo desta curva é de 0.743.

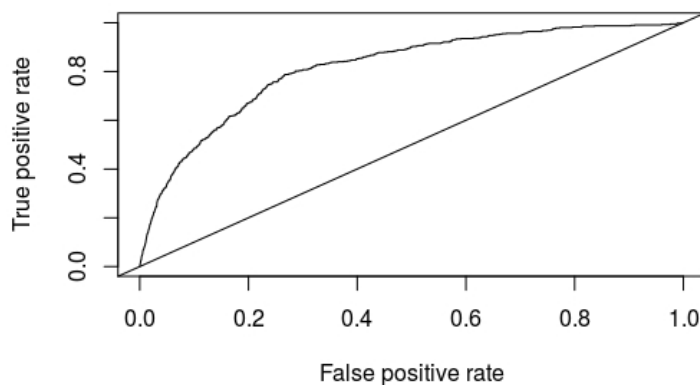


Figure 4.8: Análise da curva de ROC do modelo para as micro empresas

Apresenta-se o comportamento do modelo removendo as variáveis com p-value menos significativo, para as micro empresas.

p-value	AIC	ROC
Original	13 254.38	0.7428
]0.1; 1]	13 254.69	0.7421
]0.05; 0.1]	13 257.54	0.7427

Table 4.8: AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.9

Tendo por base o observado na tabela acima, decidiu-se optar pelo modelo original - sem remoção de variáveis -, uma vez que este apresenta os melhores valores do AIC e ROC, apesar das diferenças não serem significativas.

As taxas esperadas e observadas de execução por escalão de *rating* foram as seguintes:

Rating	Esperado	Em cumprimento	Incumprimento	Observado	Diferença
1 - Formidável	0.0059	2968	23	0.0076	-0.0018
2 - Excelente	0.0114	3098	32	0.0103	0.0012
3 - Bom	0.0165	3084	43	0.0138	0.0028
4 - Razoável	0.0196	3660	72	0.0193	0.0003
5 - Inadequado	0.0272	2985	102	0.0330	-0.0058
6 - Fraco	0.0441	3031	137	0.0432	0.0009
7 - Terrível	0.0969	2801	320	0.1025	-0.0056

Table 4.9: Escalões de rating para as micro empresas

Para o modelo das pequenas, médias e grandes empresas temos os seguintes coeficientes:

```

Coefficients:
              z value Pr(>|z|)
(Intercept) -18.717 < 2e-16 ***
VAR01        -1.353  0.176199
VAR09        -2.460  0.013911 *
VAR11         1.891  0.058675 .
VAR13         4.792  1.65e-06 ***
VAR15         4.394  1.11e-05 ***
VAR19        -3.457  0.000545 ***
VAR20        -3.651  0.000261 ***
VAR28        -2.440  0.014696 *
VAR29         1.510  0.131139
VAR34         2.989  0.002803 **
VAR39         2.754  0.005895 **
VAR40         9.087 < 2e-16 ***
VAR43         3.025  0.002489 **
VAR44        -1.494  0.135144
VAR46        -5.510  3.59e-08 ***
VAR47        -2.437  0.014804 *
VAR51         6.016  1.79e-09 ***
VAR53         4.961  7.01e-07 ***
VAR55         1.877  0.060522 .
VAR58        -1.876  0.060666 .
VAR59        -6.058  1.38e-09 ***
VAR60         3.017  0.002550 **
VAR63         4.344  1.40e-05 ***
VAR64        -6.534  6.41e-11 ***
VAR66         4.107  4.01e-05 ***
VAR68        -1.852  0.064010 .
VAR70         2.624  0.008696 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 4.9: Coeficientes do modelo das pequenas, médias e grandes empresas

A área debaixo da curva de ROC obtida para as pequenas, médias e grandes empresas é de 0.828.

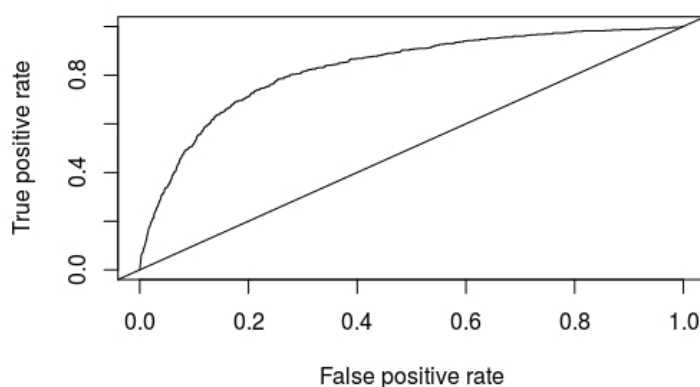


Figure 4.10: Análise da curva de ROC do modelo para as micro empresas

Apresenta-se o comportamento do modelo removendo as variáveis com p-value menos significativo, para as pequenas, medias e grandes empresas.

p-value	AIC	ROC
Original	8 494.36	0.8278
]0.1; 1]	8 494.64	0.8267
]0.05; 0.1]	8 502.41	0.8267

Table 4.10: AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.9

Tendo por base o observado na tabela acima, decidiu-se optar pelo modelo original - sem remoção de variáveis -, uma vez que este apresenta os melhores valores do AIC e ROC, apesar das diferenças não serem significativas.

As taxas, esperadas e observadas, do modelo sem remoção de variáveis, de execução por escalão de *rating* foram as seguintes:

Rating	Esperado	Em cumprimento	Incumprimento	Observado	Diferença
1 - Formidável	0.0051	1801	9	0.0050	0.0001
2 - Excelente	0.0058	1869	12	0.0064	-0.0006
3 - Bom	0.0108	1788	18	0.0100	0.0009
4 - Razoável	0.0155	2115	34	0.0158	-0.0003
5 - Inadequado	0.0307	1734	46	0.0258	0.0048
6 - Fraco	0.0567	1762	109	0.0582	-0.0015
7 - Terrível	0.1467	1533	328	0.1762	-0.0296

Table 4.11: Escalões de rating para as pequenas, médias e grandes empresas

4.2.3 Modelos construídos sem correlações superiores a 0,70

Foram ainda retiradas as variáveis cuja correlação é superior a 0,7, mantendo-se do mesmo modo a variável com melhor AIC no modelo de regressão logística univariado. Retiraram-se as 14 variáveis seguintes: VAR02, VAR03, VAR15, VAR17, VAR20, VAR25, VAR30, VAR44, VAR51, VAR60, VAR62, VAR63, VAR64 e VAR66.

Depois de retiradas estas variáveis, ficaram apenas 30 das 72 variáveis iniciais, com as quais se irá desenvolver um modelo de regressão logística que melhor se ajuste aos dados presentes no ficheiro cedido. À semelhança do que foi feito anteriormente, foi construído um modelo para as micro empresas e para as pequenas, médias e grandes empresas.

Para o modelo das micro empresas temos os seguintes coeficientes:


```

Coefficients:
              z value Pr(>|z|)
(Intercept) -15.836 < 2e-16 ***
VAR01        -1.783 0.074562 .
VAR07        -1.940 0.052432 .
VAR09        -5.050 4.42e-07 ***
VAR19        -2.664 0.007727 **
VAR28        -2.541 0.011042 *
VAR29         3.789 0.000151 ***
VAR32        -3.064 0.002183 **
VAR34         2.048 0.040590 *
VAR37         2.698 0.006986 **
VAR39        -4.419 9.90e-06 ***
VAR40         9.329 < 2e-16 ***
VAR43         5.540 3.02e-08 ***
VAR46        -4.743 2.11e-06 ***
VAR47        -1.709 0.087466 .
VAR48        -3.911 9.18e-05 ***
VAR49        -5.652 1.59e-08 ***
VAR52         4.167 3.09e-05 ***
VAR55        -4.662 3.13e-06 ***
VAR57         3.342 0.000832 ***
VAR68        -2.766 0.005676 **
VAR70         3.146 0.001653 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 4.11: Coeficientes do modelo das micro empresas

A seguir apresenta-se a curva de ROC para este modelo. A área por baixo desta curva é de 0.736.

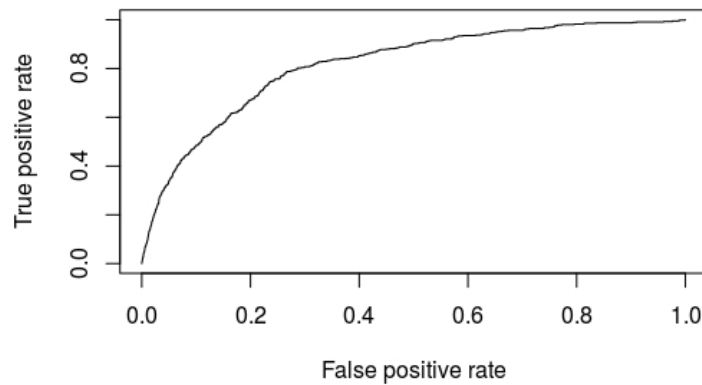


Figure 4.12: Análise da curva de ROC do modelo para as micro empresas

Analogamente ao efetuado para os modelos anteriores, apresenta-se o comportamento do modelo removendo as variáveis com p-value menos significativo.

p-value	AIC	ROC
Original	13 354.17	0.7358
]0.05; 0.1]	13 358.31	0.7361

Table 4.12: AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.7

Tendo por base o observado nesta tabela, optou-se pelo modelo original - sem remoção de variáveis -, apesar deste não apresentar o melhor valor de ROC, a diferença obtida entre os modelos é pouco significativa.

As taxas, esperadas e observadas, de execução, do modelo original, por escalão de *rating* foram as seguintes:

Rating	Esperado	Em cumprimento	Incumprimento	Observado	Diferença
1 - Formidável	0.0082	2975	24	0.0080	0.0003
2 - Excelente	0.0111	3031	34	0.0111	0.0001
3 - Bom	0.0150	3179	53	0.0163	-0.0014
4 - Razoável	0.0214	3501	66	0.0185	0.0029
5 - Inadequado	0.0273	3113	106	0.0329	-0.0057
6 - Fraco	0.0423	2990	135	0.0432	-0.0009
7 - Terrível	0.0962	2838	311	0.0988	-0.0025

Table 4.13: Escalões de rating para as micro empresas

Para o modelo das pequenas, médias e grandes empresas temos os seguintes coeficientes:

```

Coefficients:
              z value Pr(>|z|)
(Intercept) -18.550 < 2e-16 ***
VAR09       -1.421 0.155307
VAR11       -2.426 0.015251 *
VAR13        3.308 0.000938 ***
VAR19       -3.409 0.000653 ***
VAR28       -6.683 2.34e-11 ***
VAR29        2.211 0.027040 *
VAR32       -8.999 < 2e-16 ***
VAR34        2.342 0.019161 *
VAR36        1.588 0.112315
VAR40        7.663 1.81e-14 ***
VAR43        4.792 1.65e-06 ***
VAR46       -6.347 2.19e-10 ***
VAR47       -2.051 0.040264 *
VAR48       -3.502 0.000462 ***
VAR49       -2.560 0.010478 *
VAR53        5.656 1.55e-08 ***
VAR54       -1.716 0.086085 .
VAR59       -5.499 3.83e-08 ***
VAR70        2.572 0.010109 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 4.13: Coeficientes do modelo das pequenas, médias e grandes empresas

A área debaixo da curva de ROC obtida para este modelo é de 0.816. Apresenta-se a seguir gráfico desta curva:

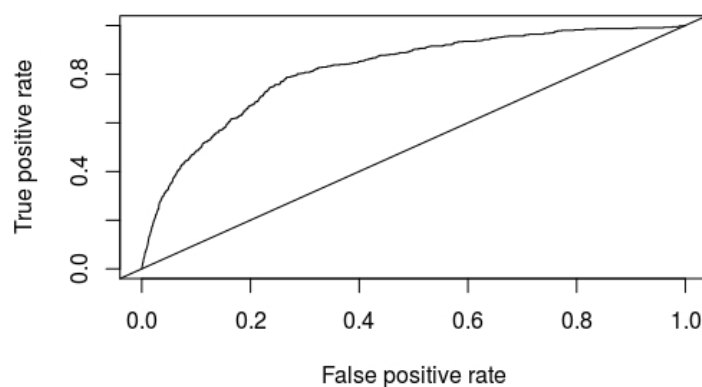


Figure 4.14: Análise da curva de ROC do modelo para as micro empresas

De seguida, apresenta-se as alterações ao AIC e ao ROC, quase se eliminam do modelo as variáveis com p-value menos significativo.

p-value	AIC	ROC
Original	8 641.42	0.8165
]0.1; 1]	8 642.95	0.8141
]0.05; 0.1]	8 644.45	0.8133

Table 4.14: AIC e ROC dos modelos com as variáveis com restrições de correlações a 0.7

Tendo por base o observado na tabela acima, decidiu-se optar pelo modelo original - sem remoção de variáveis -, uma vez que este apresenta os melhores valores do AIC e ROC, apesar das diferenças não serem significativas.

As taxas, esperadas e observadas, deste modelo, de execução por escalão de *rating* foram as seguintes:

Rating	Esperado	Em cumprimento	Incumprimento	Observado	Diferença
1 - Formidável	0.0053	1787	7	0.0039	0.0014
2 - Excelente	0.0071	1830	13	0.0070	0.0001
3 - Bom	0.0124	1883	24	0.0125	-0.0001
4 - Razoável	0.0191	2051	38	0.0181	0.0010
5 - Inadequado	0.0267	1751	43	0.0239	0.0028
6 - Fraco	0.0578	1761	138	0.0726	-0.0148
7 - Terrível	0.1420	1539	293	0.1599	-0.0179

Table 4.15: Escalões de rating para as pequenas, médias e grandes empresas

Capítulo 5

Conclusão

Para a concessão de crédito a uma empresa não financeira, é fundamental que a sua análise tenha em conta as melhores práticas bancárias.

Sendo inevitável que qualquer sistema de predição tenha limitações, os modelos apresentados denotam uma boa segmentação das empresas tendo em conta o incumprimento.

O objetivo desta dissertação é formular um rating com capacidade discriminante para, posteriormente, este procedimento, em R, ser usado na empresa que forneceu os dados. Note-se que modelo mostra ser uma ferramenta eficiente na previsão da probabilidade de incumprimento de uma empresa e identificou-se claramente as principais variáveis a ter em conta neste rating.

Os resultados obtidos, através da ferramenta de trabalho R, permitiram o escalonamento, com o uso de percentis, de cada empresa, obtendo assim um rating final.

Num contexto de orientações futuras, sugere-se uma abordagem com uma base de dados maior, relativamente ao nível temporal, visto os dados desta amostra serem somente de uma altura de crise, faltando nesta tese dados em alturas expansionistas, isto no sentido de aprimorar a fiabilidade do modelo.

Apresentam-se a seguir as conclusões dos modelos estudados, em que:

- Modelo 1 é o modelo construído a partir da base de dados completa, isto é, com todas as variáveis excepto as que se desconsideraram por motivos puramente económicos.
- Modelo 2 é o modelo construído a partir da base de dados anterior sem as variáveis cuja correlação é superior a 0,90, deixando apenas a melhor das variáveis utilizando o AIC da regressão logística univariada.
- Modelo 3 é o modelo construído a partir da base de dados considerada no modelo 2 sem as variáveis cuja correlação é superior a 0,70, deixando apenas a melhor das variáveis utilizando o AIC da regressão logística univariada.

Relativamente às micro empresas:

Rating	Modelo 1		Modelo 2		Modelo 3	
	Esperado	Diferença	Esperado	Diferença	Esperado	Diferença
1 - Formidável	0.0063	-0.0017	0.0059	-0.0018	0.0082	0.0003
2 - Excelente	0.0101	0.0021	0.0114	0.0012	0.0111	0.0001
3 - Bom	0.0165	-0.0008	0.0165	0.0028	0.0150	-0.0014
4 - Razoável	0.0189	-0.0003	0.0196	0.0003	0.0214	0.0029
5 - Inadequado	0.0279	-0.0023	0.0272	-0.0058	0.0273	-0.0057
6 - Fraco	0.0421	-0.0027	0.0441	0.0009	0.0423	-0.0009
7 - Terrível	0.0999	-0.0016	0.0969	-0.0056	0.0962	-0.0025
ROC	0.744		0.743		0.736	

Table 5.1: Análise comparativa dos modelos para as micro empresas

Nos modelos construídos para as micro empresas, note-se que as variáveis retiradas, por motivos de correlação, não influenciam o valor do ROC significativamente. A diferença do incumprimento observado para o esperado é também pouco significativa e, por isso, pode-se dizer que os modelos são ajustados para aquilo que se espera de um modelo de rating.

As variáveis explicativas de cada um dos modelos são:

Variável	Modelos	Variável	Modelos	Variável	Modelos
VAR01	3	VAR03	1 2	VAR07	1 2 3
VAR02	2	VAR09	2 3	VAR19	1 2 3
VAR04	1	VAR13	1 2	VAR29	1 2 3
VAR05	1	VAR17	1 2	VAR34	1 2 3
VAR15	2	VAR30	1 2	VAR39	1 2 3
VAR24	1	VAR37	2 3	VAR40	1 2 3
VAR27	1	VAR44	1 2	VAR43	1 2 3
VAR28	3	VAR51	1 2	VAR46	1 2 3
VAR32	3	VAR59	1 2	VAR47	1 2 3
VAR33	1	VAR63	1 2	VAR48	1 2 3
VAR36	2			VAR49	1 2 3
VAR57	3			VAR52	1 2 3
VAR64	1			VAR55	1 2 3
				VAR68	1 2 3
				VAR70	1 2 3

Table 5.2: Variáveis explicativas dos modelos para as micro empresas

Como a seleção de variáveis em cada modelo não teve em conta os anteriores e as variáveis

selecionadas por estes, apenas tem em consideração o valor de AIC da regressão logística univariada, o modelos selecionam variáveis distintas mas com correlação elevada com outras que foram removidas.

Note-se que:

- No modelo 2, foi selecionada a VAR02, contudo a mesma foi retirada devido à sua correlação com a VAR01, posteriormente, escolhida pelo algoritmo no modelo 3.
- Nos modelos 1 e 2, foram selecionadas as variáveis VAR03 e VAR17, que, devido à correlação com a VAR09 e VAR48, respetivamente, foram removidas da seleção inicial para o modelo 3.
- A VAR57 não tem correlações fortes ou muito fortes com nenhuma outra variável, sendo apenas selecionada para o modelo 3.

Relativamente às pequenas, médias e grandes empresas:

Rating	Modelo 1		Modelo 2		Modelo 3	
	Esperado	Diferença	Esperado	Diferença	Esperado	Diferença
1 - Formidável	0.0046	-0.0014	0.0051	0.0001	0.0053	0.0014
2 - Excelente	0.0057	0.0003	0.0058	-0.0006	0.0071	0.0001
3 - Bom	0.0110	0.0012	0.0108	0.0009	0.0124	-0.0001
4 - Razoável	0.0177	0.0020	0.0155	-0.0003	0.0191	0.0010
5 - Inadequado	0.0290	0.0037	0.0307	0.0048	0.0267	0.0028
6 - Fraco	0.0539	-0.0066	0.0567	-0.0015	0.0578	-0.0148
7 - Terrível	0.1487	-0.0255	0.1467	-0.0296	0.1420	-0.0179
ROC	0.826		0.828		0.816	

Table 5.3: Análise comparativa dos modelos para as pequenas, médias e grandes empresas

Nos modelos das pequenas, médias e grandes empresas, note-se que as variáveis retiradas, por motivos de correlação, não influenciam o valor do ROC significativamente. Acontece até, no segundo caso (em que foram retiradas apenas as variáveis cuja correlação é superior a 0,90), o valor do ROC ser o melhor de entre os 3 modelos.

Tal como para as micro empresas, a diferença do incumprimento observado para o esperado é também pouco significativa e, por isso, pode-se dizer que os modelos são ajustados para aquilo que se espera de um modelo de rating.

As variáveis explicativas escolhidas para cada um dos modelos são:

Variável	Modelos	Variável	Modelos	Variável	Modelos
VAR05	1	VAR01	1 2	VAR09	1 2 3
VAR07	1	VAR11	2 3	VAR13	1 2 3
VAR21	1	VAR15	1 2	VAR19	1 2 3
VAR24	1	VAR20	1 2	VAR28	1 2 3
VAR31	1	VAR29	2 3	VAR34	1 2 3
VAR32	3	VAR39	1 2	VAR40	1 2 3
VAR36	3	VAR48	1 3	VAR43	1 2 3
VAR37	1	VAR51	1 2	VAR46	1 2 3
VAR44	2	VAR54	1 3	VAR47	1 2 3
VAR49	3	VAR55	1 2	VAR53	1 2 3
VAR58	2	VAR60	1 2	VAR59	1 2 3
		VAR63	1 2	VAR70	1 2 3
		VAR64	1 2		
		VAR66	1 2		
		VAR68	1 2		

Table 5.4: Variáveis explicativas dos modelos para as pequenas, médias e grandes empresas

Em relação às variáveis selecionadas para cada modelo, note-se que:

- No modelo 1, foi selecionada a VAR05, contudo a mesma foi retirada devido à sua correlação com a VAR09.
- Nos modelos 1 e 2, foram selecionadas as variáveis VAR15 e VAR20, que, devido à correlação com a VAR48 e VAR11, respetivamente, foi removida da seleção inicial para o modelo 3.
- A VAR58 não tem correlações fortes ou muito fortes com nenhuma outra variável, sendo apenas selecionada para o modelo 2.

Como conclusão final, pode-se assim afirmar que os resultados obtidos nos testes realizados, determinam que o modelo é adequado para responder ao problema formulado, demonstrando uma capacidade prognosticadora ajustada para a sua aplicação na predição de insolvência das micro empresas e pequenas medias e grandes empresas.

Capítulo 6

Desenvolvimentos futuros

Conforme é referido na instrução consolidada 4/2011 do Banco de Portugal (ver [3]), os testes de esforço constituem ferramentas de gestão de risco utilizadas no âmbito da avaliação e gestão de risco das instituições, cuja utilidade consiste num melhor entendimento do seu perfil de risco. Em particular, os testes de esforço devem desempenhar um papel de relevo no planeamento do capital interno e da liquidez, de modo a assegurar a capacidade das instituições para absorver choques adversos. O mesmo define que um teste de esforço é a técnica de gestão de risco que visa avaliar os efeitos potenciais, nas condições financeiras de uma instituição, resultantes de alterações nos fatores de risco em função de acontecimentos excecionais, mas plausíveis.

Para a conceção e implementação dos testes de esforço, devem ser tomados em consideração as características, a dimensão e o nível de complexidade das instituições, bem como a respetiva natureza, os riscos inerentes às atividades que desenvolvem e a política de gestão desses riscos.

Será objetivo futuro elaborar um modelo de teste de esforço para o rating, calibração e elaboração de propostas tanto para definição da magnitude dos choques, como do horizonte temporal subjacente.

Bibliography

- [1] Akaike, H.; *A New Look at the Statistical Model Identification*; IEEE Transactions on Automatic Control, Vol. !9, Número 6 (dezembro de 1974), pg.716-723
- [2] Banco de Portugal, *Instrução 5/2011*, podem ser consultadas em https://www.bportugal.pt/pt-PT/sibap/Lists/SIBAPInstrucoes/Attachments/86/H_000_5_2011i.pdf
- [3] Banco de Portugal, *Instrução 4/2011*, podem ser consultadas em <http://www.bportugal.pt/sibap/application/app1/docs1/historico/textos/4-2011i.pdf>
- [4] Banco de Portugal, *Instrução 16/2013*, podem ser consultadas em <http://www.bportugal.pt/sibap/application/app1/docs1/historico/textos/16-2013i.pdf>
- [5] Banco de Portugal, *Instrução para o controlo de risco*, podem ser consultadas em https://www.bportugal.pt/pt-PT/0BancoeoEurosistema/ConsultasPublicas/Documents/Consulta_BP_2_07_MAR.pdf
- [6] David Durand *Risk Elements in Consumer Instalment Financing*, pode ser consultada em <http://www.nber.org/chapters/c9265.pdf>
- [7] Draper, N. R., e Smith, H.; *Applied Regression Analysis*; Wiley Series in Probability and Statistics; New York, Chichester, Weinheim; 3ª edição; 1998.
- [8] Harrell, E.; *Regression Modeling Strategies*; Springer Series in Statistics; New York; 1ª edição; 2001.
- [9] Hosmer, D. W., e Lemeshow, S.; *Applied Logistic Regression*; Wiley Series in Probability and Statistics; New York, Chichester, Weinheim; 2ª edição; 2000.
- [10] Informa D&B *Barómetros de empresas*, <http://biblioteca.informadb.pt/>
- [11] Instituto Português de Estatística, *Classificação Portuguesa de Actividades Económicas*, pode ser consultada em https://www.ine.pt/ine_novidades/semin/cae/CAE_REV_3.pdf
- [12] Kohavi, R.; *A study of cross-validation and bootstrap for accuracy estimation and model selection*, International Joint Conference on Artificial Intelligence; 1995, pg. 1137-1143.

- [13] Mukaka, M.M.; *A guide to appropriate use of Correlation coefficient in medical research*, Malawi Med J; 2012, pg. 69-71.
- [14] Poliaprende *Gestão financeira*, http://poliemprende.ipvc.pt/files/13.Gest__o_financeira.pdf
- [15] Abdou, Hussein, e Pointon, John; *Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature*, Intelligent Systems in Accounting, Finance & Management, 18 (2-3); 2011, pg. 59-88.
- [16] *R-Project*, cujo manual de instalação e manuais de ajuda podem ser consultados em <https://cran.r-project.org/>
- [17] Swets, J. A.; *The Relative Operating Characteristic in Psychology*; Science, New Series, Volume 182, Publicação 4116 (1973/12/07), pg. 990-1000

Appendix A

Tabela de correlações de todas as variáveis contabilísticas

Variavel	VAR01	VAR02	VAR03	VAR04	VAR05	VAR07	VAR08	VAR09	VAR10	VAR11	VAR13
VAR01	1	0.77	0.55	0.42	0.64	0.59	0.28	0.24	0.24	0.48	0.56
VAR02	0.77	1	0.64	0.46	0.66	0.64	0.31	0.27	0.27	0.63	0.47
VAR03	0.55	0.64	1	0.95	0.45	0.41	0.88	0.85	0.85	0.36	0.34
VAR04	0.42	0.46	0.95	1	0.3	0.25	0.97	0.94	0.94	0.13	0.25
VAR05	0.64	0.66	0.45	0.3	1	0.9	0.08	0.04	0.04	0.64	0.48
VAR07	0.59	0.64	0.41	0.25	0.9	1	0	-0.03	-0.03	0.65	0.47
VAR08	0.28	0.31	0.88	0.97	0.08	0	0	-0.03	-0.03	0.65	0.47
VAR09	0.24	0.27	0.85	0.94	0.04	-0.03	1	0.98	0.98	-0.03	0.14
VAR10	0.24	0.27	0.85	0.94	0.04	-0.03	0.98	1	1	-0.03	0.11
VAR11	0.48	0.63	0.36	0.13	0.64	0.65	0.98	1	1	-0.03	0.11
VAR13	0.56	0.47	0.34	0.25	0.48	0.47	-0.03	-0.03	-0.03	1	0.36
VAR15	0.48	0.53	0.46	0.39	0.73	0.61	0.14	0.11	0.11	0.36	1
VAR17	0.73	0.72	0.63	0.52	0.83	0.7	0.14	0.11	0.11	0.36	1
VAR19	0.57	0.68	0.64	0.52	0.64	0.56	0.24	0.21	0.21	0.38	0.26
VAR20	0.43	0.53	0.22	0.04	0.72	0.74	0.24	0.21	0.21	0.38	0.26
VAR21	0.47	0.57	0.3	0.11	0.78	0.77	0.36	0.3	0.3	0.44	0.55
VAR23	0.52	0.56	0.44	0.34	0.74	0.6	0.36	0.3	0.3	0.44	0.55
VAR24	0.69	0.7	0.58	0.46	0.88	0.77	0.39	0.38	0.38	0.63	0.47
VAR25	0.67	0.72	0.53	0.38	0.93	0.84	0.39	0.38	0.38	0.63	0.47
VAR27	0.99	0.78	0.59	0.46	0.69	0.63	-0.15	-0.16	-0.16	0.76	0.4
VAR28	-0.03	-0.02	-0.01	0	-0.05	-0.05	-0.08	-0.1	-0.1	0.8	0.41
VAR29	0	0.04	0.02	0	0.11	0.13	0.19	0.15	0.15	0.42	0.33
VAR30	-0.01	0	0.02	0.03	-0.02	-0.02	0.19	0.15	0.15	0.42	0.33
VAR31	-0.02	-0.01	0.02	0.02	-0.04	-0.05	0.28	0.23	0.23	0.54	0.47
VAR32	0	0.01	0.03	0.03	-0.03	-0.03	0.17	0.13	0.13	0.69	0.49
VAR33	0	0	0.02	0.02	-0.01	-0.01	0.32	0.27	0.27	0.49	0.56
VAR34	0	0.02	0.02	0.02	-0.04	-0.05	0.32	0.27	0.27	0.49	0.56
VAR36	-0.07	-0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	-0.03	-0.06
VAR37	-0.01	-0.02	-0.01	0	0	0.01	-0.03	-0.04	-0.04	0.09	0.09
VAR38	0.09	0.03	0	0.01	-0.05	-0.05	0.04	0.03	0.03	-0.04	-0.06
VAR39	0.09	0.03	0	0.01	-0.05	-0.05	0.04	0.04	0.04	-0.02	-0.01
VAR40	-0.03	0.02	0.01	-0.01	0.08	0.09	0.04	0.05	0.05	0.01	0
VAR43	0.02	0.01	0	0.01	0.05	0.05	0.03	0.02	0.02	-0.04	-0.07
VAR44	-0.08	0	0.04	0.03	0.02	0.02	0.04	0.03	0.03	-0.06	-0.13
VAR45	0.24	0.27	0.85	0.94	0.04	-0.03	0.01	0.01	0.01	0.02	-0.02
VAR46	-0.09	-0.05	-0.01	0.01	-0.05	-0.06	0.01	0.01	0.01	0.02	-0.02
VAR47	-0.07	-0.06	-0.02	-0.01	-0.02	-0.02	0	-0.01	-0.01	-0.04	-0.06
VAR48	-0.65	-0.54	-0.42	-0.33	-0.71	-0.64	0.02	0.02	0.02	-0.08	-0.06
VAR49	-0.09	-0.09	-0.03	-0.01	-0.04	-0.05	0.03	0.03	0.03	-0.08	-0.06
VAR51	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.04	-0.04	-0.04	0.1	0.08
VAR52	-0.04	0	0.01	0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	0.01
VAR53	0.04	0	-0.01	0.01	-0.01	-0.02	0.02	0.02	0.02	0.05	-0.03
VAR54	-0.04	-0.02	0.01	0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	0.01
VAR55	0.03	0.08	0.04	0.01	0.05	0.05	0.02	0.02	0.02	0.05	-0.03
VAR57	0.06	0	-0.01	0.01	0	-0.01	0.98	1	1	-0.03	0.11
VAR58	-0.02	-0.04	-0.02	0.01	-0.05	-0.07	0.02	0.03	0.03	-0.05	-0.13
VAR59	0.13	0.02	0.03	0.03	0.01	0.01	-0.01	-0.01	-0.01	-0.02	-0.04
VAR60	0.06	0.04	0	-0.01	0.1	0.12	-0.17	-0.13	-0.13	-0.35	-0.4
VAR62	-0.01	0.01	0.01	0	-0.01	0	0	0.01	0.01	-0.05	0.07
VAR63	0.11	-0.02	0	0.01	0.01	0.02	-0.01	-0.02	-0.02	-0.03	-0.07
VAR64	0.02	0.04	0.03	0.01	0.07	0.08	-0.01	-0.02	-0.02	-0.03	-0.07
VAR66	0.02	0	-0.03	-0.03	0.03	0.03	0.02	0.03	0.03	-0.02	-0.08
VAR67	0.05	0.04	0.01	-0.01	0.06	0.07	0.02	0.02	0.02	-0.1	0.02
VAR68	0.04	0.04	0.01	-0.01	0.05	0.06	0.02	0.02	0.02	-0.03	0
VAR70	-0.17	-0.17	-0.1	-0.07	-0.09	-0.1	0	0	0	0.09	-0.12

Variavel	VAR15	VAR17	VAR19	VAR20	VAR21	VAR23	VAR24	VAR25	VAR27	VAR28	VAR29
VAR01	0.48	0.73	0.57	0.43	0.47	0.52	0.69	0.67	0.99	-0.03	0
VAR02	0.53	0.72	0.68	0.53	0.57	0.56	0.7	0.72	0.78	-0.02	0.04
VAR03	0.46	0.63	0.64	0.22	0.3	0.44	0.58	0.53	0.59	-0.01	0.02
VAR04	0.39	0.52	0.52	0.04	0.11	0.34	0.46	0.38	0.46	0	0
VAR05	0.73	0.83	0.64	0.72	0.78	0.74	0.88	0.93	0.69	-0.05	0.11
VAR07	0.61	0.7	0.56	0.74	0.77	0.6	0.77	0.84	0.63	-0.05	0.13
VAR08	0.24	0.36	0.39	-0.15	-0.08	0.19	0.77	0.84	0.63	-0.05	0.13
VAR09	0.21	0.3	0.38	-0.16	-0.1	0.15	0.28	0.17	0.32	0.01	-0.03
VAR10	0.21	0.3	0.38	-0.16	-0.1	0.15	0.23	0.13	0.27	0.01	-0.04
VAR11	0.38	0.44	0.63	0.76	0.8	0.42	0.23	0.13	0.27	0.01	-0.04
VAR13	0.26	0.55	0.47	0.4	0.41	0.33	0.54	0.69	0.49	-0.03	0.09
VAR15	1	0.83	0.56	0.43	0.53	0.84	0.47	0.49	0.56	-0.06	0.09
VAR17	0.83	1	0.69	0.49	0.58	0.8	0.47	0.49	0.56	-0.06	0.09
VAR19	0.56	0.69	1	0.52	0.58	0.55	0.83	0.8	0.55	-0.02	0.04
VAR20	0.43	0.49	0.52	1	0.96	0.46	0.83	0.8	0.55	-0.02	0.04
VAR21	0.53	0.58	0.58	0.96	1	0.54	0.94	0.9	0.78	-0.03	0.05
VAR23	0.84	0.8	0.55	0.46	0.54	1	0.94	0.9	0.78	-0.03	0.05
VAR24	0.83	0.94	0.65	0.54	0.63	0.83	0.65	0.68	0.61	-0.03	0.06
VAR25	0.8	0.9	0.68	0.74	0.83	0.8	0.65	0.68	0.61	-0.03	0.06
VAR27	0.55	0.78	0.61	0.46	0.51	0.58	0.54	0.74	0.46	-0.04	0.11
VAR28	-0.02	-0.03	-0.03	-0.04	-0.04	-0.02	0.63	0.83	0.51	-0.04	0.11
VAR29	0.04	0.05	0.06	0.11	0.11	0.04	0.83	0.8	0.58	-0.02	0.04
VAR30	-0.01	-0.02	0	-0.04	-0.04	-0.02	0.83	0.8	0.58	-0.02	0.04
VAR31	-0.01	-0.01	0.06	-0.04	-0.04	-0.02	1	0.96	0.75	-0.03	0.05
VAR32	0	0	0.09	-0.03	-0.03	-0.02	0.96	1	0.73	-0.04	0.08
VAR33	-0.01	-0.02	-0.03	-0.03	-0.03	-0.01	0.75	0.73	1	-0.03	0
VAR34	0	-0.02	-0.05	-0.06	-0.06	-0.01	0.75	0.73	1	-0.03	0
VAR36	0.01	-0.01	0.02	0.04	0.04	0.01	-0.03	-0.04	-0.03	1	-0.29
VAR37	-0.01	-0.03	-0.08	-0.01	-0.01	-0.01	0.05	0.08	0	-0.29	1
VAR38	-0.03	-0.02	-0.07	-0.08	-0.08	-0.04	-0.03	-0.04	-0.01	0.11	-0.08
VAR39	-0.03	-0.02	-0.07	-0.08	-0.08	-0.03	-0.04	-0.04	-0.02	0.1	-0.18
VAR40	0.03	0.03	0.06	0.13	0.12	0.02	-0.02	-0.03	0	0.1	-0.18
VAR43	0.02	0.02	0	-0.02	-0.01	0.02	-0.02	-0.03	0	0.07	0
VAR44	0.01	-0.01	0.04	0.05	0.04	0.02	-0.02	-0.04	0	0.17	-0.17
VAR45	0.21	0.3	0.38	-0.16	-0.1	0.15	-0.01	0.01	-0.07	0.02	0.3
VAR46	-0.01	-0.05	-0.01	-0.04	-0.04	-0.02	-0.01	0.01	-0.07	0.02	0.3
VAR47	0	-0.02	-0.01	-0.01	-0.01	0	-0.01	-0.01	-0.01	0	0.1
VAR48	-0.71	-0.78	-0.37	-0.38	-0.48	-0.68	-0.02	-0.05	0.08	0.05	-0.43
VAR49	-0.05	-0.04	-0.01	-0.03	-0.03	-0.03	-0.02	-0.05	0.08	0.05	-0.43
VAR51	-0.01	-0.03	-0.06	-0.01	-0.01	0	0.04	0.07	-0.03	-0.14	0.59
VAR52	0.01	0	0.02	-0.01	-0.01	0	0.02	0.01	0.02	-0.09	0.33
VAR53	0.02	0.04	-0.02	-0.05	-0.05	0	-0.01	0.01	-0.07	0.13	0.27
VAR54	0	0	0.01	-0.01	-0.01	-0.01	0.02	0.01	0.02	-0.09	0.33
VAR55	0.05	0.01	0.05	0.06	0.06	0.04	-0.01	0.01	-0.07	0.13	0.27
VAR57	0.02	0.05	-0.02	-0.04	-0.04	0	0.23	0.13	0.27	0.01	-0.04
VAR58	-0.01	0.01	-0.01	-0.09	-0.08	-0.03	-0.06	-0.06	-0.08	0.08	-0.09
VAR59	-0.02	0.03	0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.07	0.01	0.07
VAR60	0.04	0.06	0.04	0.09	0.08	0.03	-0.82	-0.77	-0.69	0.02	-0.03
VAR62	-0.02	-0.04	-0.09	-0.01	-0.01	-0.01	-0.06	-0.05	-0.09	0.02	-0.02
VAR63	-0.02	0.04	0.01	0	-0.01	-0.01	-0.01	-0.01	-0.02	0.01	-0.05
VAR64	0.01	0.03	0.05	0.12	0.1	-0.02	-0.01	-0.01	-0.02	0.01	-0.05
VAR66	0	0	-0.09	0.03	0.03	0.02	-0.02	-0.02	-0.03	0.05	-0.01
VAR67	0.02	0.04	0.03	0.07	0.07	0.01	0.01	-0.01	0.03	-0.01	-0.11
VAR68	0.02	0.03	0.02	0.06	0.05	0.01	-0.02	-0.02	-0.04	0.03	0.02
VAR70	-0.04	-0.12	-0.12	-0.09	-0.08	-0.06	0.04	0.05	0.03	0	0.15

Variavel	VAR30	VAR31	VAR32	VAR33	VAR34	VAR36	VAR37	VAR38	VAR39	VAR40	VAR43
VAR01	-0.01	-0.02	0	0	0	-0.07	-0.01	0.09	0.09	-0.03	0.02
VAR02	0	-0.01	0.01	0	0.02	-0.02	-0.02	0.03	0.03	0.02	0.01
VAR03	0.02	0.02	0.03	0.02	0.02	0.02	-0.01	0	0	0.01	0
VAR04	0.03	0.02	0.03	0.02	0.02	0.02	0	0.01	0.01	-0.01	0.01
VAR05	-0.02	-0.04	-0.03	-0.01	-0.04	0.02	0	-0.05	-0.05	0.08	0.05
VAR07	-0.02	-0.05	-0.03	-0.01	-0.05	0.02	0.01	-0.05	-0.05	0.09	0.05
VAR08	0.04	0.04	0.04	0.03	0.04	0.01	0.01	-0.05	-0.05	0.09	0.05
VAR09	0.03	0.04	0.05	0.02	0.03	0.01	0	0.02	0.03	-0.04	-0.01
VAR10	0.03	0.04	0.05	0.02	0.03	0.01	-0.01	0.02	0.03	-0.04	-0.01
VAR11	-0.04	-0.02	0.01	-0.04	-0.06	0.02	-0.01	0.02	0.03	-0.04	-0.01
VAR13	-0.06	-0.01	0	-0.07	-0.13	-0.02	-0.04	-0.08	-0.08	0.1	-0.01
VAR15	-0.01	-0.01	0	-0.01	0	0.01	-0.06	-0.06	-0.06	0.08	0.01
VAR17	-0.02	-0.01	0	-0.02	-0.02	-0.01	-0.06	-0.06	-0.06	0.08	0.01
VAR19	0	0.06	0.09	-0.03	-0.05	0.02	-0.01	-0.03	-0.03	0.03	0.02
VAR20	-0.04	-0.04	-0.03	-0.03	-0.06	0.04	-0.01	-0.03	-0.03	0.03	0.02
VAR21	-0.04	-0.04	-0.03	-0.03	-0.06	0.04	-0.03	-0.02	-0.02	0.03	0.02
VAR23	-0.02	-0.02	-0.02	-0.01	-0.01	0.01	-0.03	-0.02	-0.02	0.03	0.02
VAR24	-0.03	-0.04	-0.02	-0.02	-0.02	-0.01	-0.08	-0.07	-0.07	0.06	0
VAR25	-0.04	-0.04	-0.03	-0.03	-0.04	0.01	-0.08	-0.07	-0.07	0.06	0
VAR27	-0.01	-0.02	0	0	0	-0.07	-0.01	-0.08	-0.08	0.13	-0.02
VAR28	0.11	0.1	0.1	0.07	0.17	0.02	-0.01	-0.08	-0.08	0.12	-0.01
VAR29	-0.08	-0.18	-0.18	0	-0.17	0.3	-0.01	-0.04	-0.03	0.02	0.02
VAR30	1	0.37	0.38	0.91	0.59	0.71	-0.01	-0.04	-0.03	0.02	0.02
VAR31	0.37	1	0.96	0.1	0.14	0.12	-0.01	-0.02	-0.02	0.04	0.02
VAR32	0.38	0.96	1	0.09	0.13	0.12	-0.01	-0.05	-0.05	0.07	0.01
VAR33	0.91	0.1	0.09	1	0.63	0.74	-0.01	0.08	0.08	-0.03	0.02
VAR34	0.59	0.14	0.13	0.63	1	0.37	-0.01	0.08	0.08	-0.03	0.02
VAR36	0.71	0.12	0.12	0.74	0.37	1	0	0.05	0.05	-0.14	-0.09
VAR37	0.54	-0.45	-0.48	0.78	0.46	0.55	0.1	-0.43	-0.43	0.59	0.33
VAR38	0.18	-0.03	-0.05	0.24	0.3	-0.38	0.54	0.18	0.18	-0.19	0.15
VAR39	0.18	-0.03	-0.05	0.24	0.3	-0.38	-0.45	-0.03	-0.03	-0.21	0.04
VAR40	-0.19	-0.21	-0.21	-0.1	-0.23	0.29	-0.48	-0.05	-0.05	-0.21	0.04
VAR43	0.15	0.04	0.04	0.12	0.05	0.06	0.78	0.24	0.24	-0.1	0.12
VAR44	0.46	0.16	0.17	0.45	0.46	0.75	0.46	0.3	0.3	-0.23	0.05
VAR45	0.03	0.04	0.05	0.02	0.03	0.01	0.55	-0.38	-0.38	0.29	0.06
VAR46	0.3	0.34	0.34	0.17	0.15	0.18	0.55	-0.38	-0.38	0.29	0.06
VAR47	0	-0.01	-0.01	0.01	0	0.18	1	0.23	0.23	0.03	0.08
VAR48	0.05	0.06	0.05	0.03	0.01	0.04	0.23	1	0.99	-0.59	0.09
VAR49	0.1	0.2	0.19	0.01	-0.26	0.08	0.23	0.99	1	-0.6	0.09
VAR51	0.56	-0.26	-0.29	0.73	0.47	0.55	0.03	-0.59	-0.6	1	-0.35
VAR52	0.2	0.26	0.27	0.1	0.19	0.13	0.08	0.09	0.09	-0.35	1
VAR53	0.01	0	-0.01	0.01	0.17	-0.1	0.29	-0.44	-0.45	0.33	-0.01
VAR54	0.11	0.18	0.19	0.04	0.03	0.09	0.08	0.09	0.09	-0.35	1
VAR55	0.01	-0.03	-0.01	0.03	0.3	0.11	0.29	-0.44	-0.45	0.33	-0.01
VAR57	-0.11	-0.14	-0.16	-0.05	0.03	-0.18	-0.01	0.02	0.03	-0.04	-0.01
VAR58	0.19	0.5	0.5	0.05	0.16	-0.03	-0.05	-0.02	-0.03	-0.03	-0.05
VAR59	-0.03	-0.08	-0.08	0.01	-0.06	-0.31	0.02	-0.24	-0.25	0.16	-0.05
VAR60	-0.37	-0.28	-0.28	-0.28	-0.47	-0.15	0	-0.02	-0.02	-0.01	-0.03
VAR62	0.32	-0.5	-0.53	0.53	0.54	0.3	-0.1	-0.12	-0.12	0.07	-0.08
VAR63	-0.1	-0.11	-0.11	-0.07	-0.2	-0.24	0.81	0.17	0.17	-0.03	-0.02
VAR64	-0.01	-0.04	-0.02	0	-0.12	0.12	0.81	0.17	0.17	-0.03	-0.02
VAR66	-0.32	-0.77	-0.81	-0.03	-0.06	-0.09	-0.07	-0.05	-0.05	0.02	-0.01
VAR67	-0.37	-0.25	-0.25	-0.29	-0.35	-0.18	0.01	0.19	0.19	-0.22	0.12
VAR68	-0.31	-0.21	-0.21	-0.24	-0.28	-0.16	-0.08	-0.08	-0.08	0.06	-0.01
VAR70	0.06	0.05	0.04	0.05	0.06	0.09	0.03	-0.11	-0.11	0.14	0.01

Variavel	VAR44	VAR45	VAR46	VAR47	VAR48	VAR49	VAR51	VAR52	VAR53	VAR54	VAR55
VAR01	-0.08	0.24	-0.09	-0.07	-0.65	-0.09	-0.02	-0.04	0.04	-0.04	0.03
VAR02	0	0.27	-0.05	-0.06	-0.54	-0.09	-0.02	0	0	-0.02	0.08
VAR03	0.04	0.85	-0.01	-0.02	-0.42	-0.03	-0.02	0.01	-0.01	0.01	0.04
VAR04	0.03	0.94	0.01	-0.01	-0.33	-0.01	-0.01	0.02	0.01	0.01	0.01
VAR05	0.02	0.04	-0.05	-0.02	-0.71	-0.04	-0.01	-0.01	-0.01	-0.01	0.05
VAR07	0.02	-0.03	-0.06	-0.02	-0.64	-0.05	-0.01	-0.02	-0.02	-0.02	0.05
VAR08	0.02	0.98	0.02	-0.01	-0.17	0	-0.01	-0.02	-0.02	-0.02	0.05
VAR09	0.02	1	0.03	-0.01	-0.13	0.01	-0.01	0.02	0.02	0.02	0
VAR10	0.02	1	0.03	-0.01	-0.13	0.01	-0.02	0.03	0.02	0.02	0
VAR11	0.05	-0.03	-0.05	-0.02	-0.35	-0.05	-0.02	0.03	0.02	0.02	0
VAR13	-0.03	0.11	-0.13	-0.04	-0.4	0.07	-0.03	-0.02	-0.1	-0.03	0.09
VAR15	0.01	0.21	-0.01	0	-0.71	-0.05	-0.07	-0.08	0.02	0	-0.12
VAR17	-0.01	0.3	-0.05	-0.02	-0.78	-0.04	-0.07	-0.08	0.02	0	-0.12
VAR19	0.04	0.38	-0.01	-0.01	-0.37	-0.01	-0.01	0.01	0.02	0	0.05
VAR20	0.05	-0.16	-0.04	-0.01	-0.38	-0.03	-0.01	0.01	0.02	0	0.05
VAR21	0.04	-0.1	-0.04	-0.01	-0.48	-0.03	-0.03	0	0.04	0	0.01
VAR23	0.02	0.15	-0.02	0	-0.68	-0.03	-0.03	0	0.04	0	0.01
VAR24	-0.01	0.23	-0.06	-0.02	-0.82	-0.06	-0.06	0.02	-0.02	0.01	0.05
VAR25	0.01	0.13	-0.06	-0.02	-0.77	-0.05	-0.06	0.02	-0.02	0.01	0.05
VAR27	-0.07	0.27	-0.08	-0.07	-0.69	-0.09	-0.01	-0.01	-0.05	-0.01	0.06
VAR28	0.13	0.01	0.08	0.01	0.02	0.02	-0.01	-0.01	-0.05	-0.01	0.06
VAR29	0.27	-0.04	-0.09	0.07	-0.03	-0.02	0	0	0	-0.01	0.04
VAR30	0.46	0.03	0.3	0	0.05	0.1	0	0	0	-0.01	0.04
VAR31	0.16	0.04	0.34	-0.01	0.06	0.2	-0.01	-0.02	0.01	-0.02	0.04
VAR32	0.17	0.05	0.34	-0.01	0.05	0.19	-0.01	-0.02	-0.01	-0.02	0.05
VAR33	0.45	0.02	0.17	0.01	0.03	0.01	-0.02	-0.03	0.03	-0.04	0.03
VAR34	0.46	0.03	0.15	0	0.01	-0.26	-0.02	-0.03	0.03	-0.04	0.03
VAR36	0.75	0.01	0.18	0.18	0.04	0.08	0.01	0.05	-0.01	0.03	0
VAR37	0.29	-0.01	-0.05	0.02	0	-0.1	-0.05	-0.01	-0.11	0.02	0.15
VAR38	-0.44	0.02	-0.02	-0.24	-0.02	-0.12	0.56	0.2	0.01	0.11	0.01
VAR39	-0.45	0.03	-0.03	-0.25	-0.02	-0.12	-0.26	0.26	0	0.18	-0.03
VAR40	0.33	-0.04	-0.03	0.16	-0.01	0.07	-0.29	0.27	-0.01	0.19	-0.01
VAR43	-0.01	-0.01	-0.05	-0.05	-0.03	-0.08	0.73	0.1	0.01	0.04	0.03
VAR44	1	0.02	0.22	0.17	0.05	0.07	0.47	0.19	0.17	0.03	0.3
VAR45	0.02	1	0.03	-0.01	-0.13	0.01	0.55	0.13	-0.1	0.09	0.11
VAR46	0.22	0.03	1	0.05	0.14	0.5	0.55	0.13	-0.1	0.09	0.11
VAR47	0.17	-0.01	0.05	1	0.02	0.04	0.81	-0.07	0.01	-0.08	0.03
VAR48	0.05	-0.13	0.14	0.02	1	0.12	0.17	-0.05	0.19	-0.08	-0.11
VAR49	0.07	0.01	0.5	0.04	0.12	1	0.17	-0.05	0.19	-0.08	-0.11
VAR51	0.26	-0.02	0.02	0.09	0.01	-0.04	-0.03	0.02	-0.22	0.06	0.14
VAR52	0.17	0.03	0.66	0.04	0.08	0.18	-0.02	-0.01	0.12	-0.01	0.01
VAR53	-0.2	0.02	-0.14	0.01	-0.04	-0.19	0.26	0.17	-0.2	0.1	0.2
VAR54	0.1	0.02	0.38	0.03	0.06	0.37	-0.02	-0.01	0.12	-0.01	0.01
VAR55	0.2	0	-0.06	0.01	-0.05	-0.67	0.26	0.17	-0.2	0.1	0.2
VAR57	-0.32	0.01	-0.4	-0.02	-0.07	-0.24	-0.02	0.03	0.02	0.02	0
VAR58	-0.11	0.03	0.16	0.01	0.03	0.15	0.02	0.66	-0.14	0.38	-0.06
VAR59	-0.4	0.02	-0.27	-0.18	-0.05	-0.02	0.09	0.04	0.01	0.03	0.01
VAR60	-0.31	-0.04	-0.22	-0.03	-0.06	-0.06	0.01	0.08	-0.04	0.06	-0.05
VAR62	0.4	0	-0.05	-0.02	0	-0.14	-0.04	0.18	-0.19	0.37	-0.67
VAR63	-0.39	0	-0.35	-0.08	-0.06	0.03	1	-0.04	-0.01	-0.06	0.01
VAR64	0.14	-0.02	0.16	0.01	0.01	0.17	1	-0.04	-0.01	-0.06	0.01
VAR66	-0.12	-0.04	-0.33	0.01	-0.06	-0.21	-0.04	1	0.01	0.7	0.17
VAR67	-0.2	-0.03	-0.15	-0.03	-0.05	-0.03	-0.01	0.01	1	0	-0.15
VAR68	-0.17	-0.02	-0.12	-0.03	-0.04	-0.03	-0.06	0.7	0	1	-0.07
VAR70	0.11	-0.03	0.19	0.14	0.09	0.04	0.01	0.17	-0.15	-0.07	1

Variavel	VAR57	VAR58	VAR59	VAR60	VAR62	VAR63	VAR64	VAR66	VAR67	VAR68	VAR70
VAR01	0.06	-0.02	0.13	0.06	-0.01	0.11	0.02	0.02	0.05	0.04	-0.17
VAR02	0	-0.04	0.02	0.04	0.01	-0.02	0.04	0	0.04	0.04	-0.17
VAR03	-0.01	-0.02	0.03	0	0.01	0	0.03	-0.03	0.01	0.01	-0.1
VAR04	0.01	0.01	0.03	-0.01	0	0.01	0.01	-0.03	-0.01	-0.01	-0.07
VAR05	0	-0.05	0.01	0.1	-0.01	0.01	0.07	0.03	0.06	0.05	-0.09
VAR07	-0.01	-0.07	0.01	0.12	0	0.02	0.08	0.03	0.07	0.06	-0.1
VAR08	0.01	0.03	0.02	-0.04	0.01	0	0.08	0.03	0.07	0.06	-0.1
VAR09	0.01	0.03	0.02	-0.04	0	0	-0.02	-0.04	-0.03	-0.02	-0.04
VAR10	0.01	0.03	0.02	-0.04	0	0	-0.02	-0.04	-0.03	-0.02	-0.03
VAR11	-0.09	-0.1	0	0.06	-0.03	0	-0.02	-0.04	-0.03	-0.02	-0.03
VAR13	0.05	0.02	0.09	0.12	-0.08	0.12	0.08	-0.01	0.05	0.05	-0.09
VAR15	0.02	-0.01	-0.02	0.04	-0.02	-0.02	0.07	0	0.09	0.07	-0.19
VAR17	0.05	0.01	0.03	0.06	-0.04	0.04	0.07	0	0.09	0.07	-0.19
VAR19	-0.02	-0.01	0.01	0.04	-0.09	0.01	0.01	0	0.02	0.02	-0.04
VAR20	-0.04	-0.09	-0.01	0.09	-0.01	0	0.01	0	0.02	0.02	-0.04
VAR21	-0.04	-0.08	-0.01	0.08	-0.01	-0.01	0.03	0	0.04	0.03	-0.12
VAR23	0	-0.03	-0.01	0.03	-0.01	-0.01	0.03	0	0.04	0.03	-0.12
VAR24	0.03	-0.03	0.02	0.06	-0.01	0.02	0.05	-0.09	0.03	0.02	-0.12
VAR25	0.01	-0.05	0.01	0.07	-0.02	0.02	0.05	-0.09	0.03	0.02	-0.12
VAR27	0.05	-0.02	0.12	0.06	-0.01	0.1	0.12	0.03	0.07	0.06	-0.09
VAR28	-0.04	0.05	-0.02	-0.39	0.09	-0.03	0.1	0.03	0.07	0.05	-0.08
VAR29	-0.11	-0.21	-0.2	0.7	0.04	-0.1	-0.02	0.02	0.01	0.01	-0.06
VAR30	-0.11	0.19	-0.03	-0.37	0.32	-0.1	-0.02	0.02	0.01	0.01	-0.06
VAR31	-0.14	0.5	-0.08	-0.28	-0.5	-0.11	0.02	0.03	0.04	0.03	-0.1
VAR32	-0.16	0.5	-0.08	-0.28	-0.53	-0.11	0.05	0.03	0.05	0.05	-0.1
VAR33	-0.05	0.05	0.01	-0.28	0.53	-0.07	0.02	0.01	0.05	0.04	-0.16
VAR34	0.03	0.16	-0.06	-0.47	0.54	-0.2	0.02	0.01	0.05	0.04	-0.16
VAR36	-0.18	-0.03	-0.31	-0.15	0.3	-0.24	-0.12	-0.08	-0.08	-0.05	0.05
VAR37	0.05	-0.25	0.05	-0.08	0.81	0	0.4	0.18	0.32	0.27	-0.04
VAR38	0.22	0.12	0.49	-0.15	0.28	0.26	-0.01	-0.32	-0.37	-0.31	0.06
VAR39	0.23	0.12	0.5	-0.15	0.27	0.27	-0.04	-0.77	-0.25	-0.21	0.05
VAR40	-0.23	-0.3	-0.29	0.37	0	-0.15	-0.02	-0.81	-0.25	-0.21	0.04
VAR43	0.1	0.1	0.04	0.33	0.04	0.01	0	-0.03	-0.29	-0.24	0.05
VAR44	-0.32	-0.11	-0.4	-0.31	0.4	-0.39	-0.12	-0.06	-0.35	-0.28	0.06
VAR45	0.01	0.03	0.02	-0.04	0	0	0.12	-0.09	-0.18	-0.16	0.09
VAR46	-0.4	0.16	-0.27	-0.22	-0.05	-0.35	0.12	-0.09	-0.18	-0.16	0.09
VAR47	-0.02	0.01	-0.18	-0.03	-0.02	-0.08	0	0.39	-0.12	-0.1	0.03
VAR48	-0.07	0.03	-0.05	-0.06	0	-0.06	-0.19	0.08	-0.14	-0.11	-0.05
VAR49	-0.24	0.15	-0.02	-0.06	-0.14	0.03	-0.19	0.08	-0.14	-0.1	-0.06
VAR51	0.02	-0.16	0.01	-0.22	0.58	0	0.74	0.21	0.54	0.47	0
VAR52	-0.26	0.17	-0.22	-0.11	-0.07	-0.29	-0.37	-0.04	-0.25	-0.26	-0.01
VAR53	0.66	0.51	0.16	0.02	-0.05	0.14	0.14	-0.12	-0.2	-0.17	0.11
VAR54	-0.16	0.16	-0.11	-0.03	-0.1	-0.12	-0.37	-0.04	-0.25	-0.26	-0.01
VAR55	-0.28	-0.33	-0.3	0.03	0.09	-0.36	0.14	-0.12	-0.2	-0.17	0.11
VAR57	1	0.43	0.32	0.08	-0.03	0.36	-0.02	-0.04	-0.03	-0.02	-0.03
VAR58	0.43	1	0.1	-0.14	-0.36	0.1	0.16	-0.33	-0.15	-0.12	0.19
VAR59	0.32	0.1	1	0.05	0.03	0.72	0.01	0.01	-0.03	-0.03	0.14
VAR60	0.08	-0.14	0.05	1	-0.21	0.12	0.01	-0.06	-0.05	-0.04	0.09
VAR62	-0.03	-0.36	0.03	-0.21	1	-0.07	0.17	-0.21	-0.03	-0.03	0.04
VAR63	0.36	0.1	0.72	0.12	-0.07	1	-0.05	0.22	-0.22	-0.17	0.07
VAR64	-0.3	-0.2	-0.09	0.31	0.01	-0.06	-0.05	0.22	-0.22	-0.17	0.07
VAR66	0.16	-0.37	0.08	0.25	0.43	0.11	0.14	-0.25	-0.07	-0.06	0.09
VAR67	0.01	-0.17	0.04	0.47	-0.13	0.1	-0.22	0.02	-0.04	-0.04	-0.09
VAR68	0	-0.16	0.03	0.38	-0.1	0.07	0.13	-0.18	-0.02	-0.02	0.02
VAR70	-0.15	-0.03	-0.28	-0.1	0.03	-0.24	0.09	0.03	0.03	0.03	0.1